Quan Zhang · Hong Yang *Editors*

Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings

Rasch and the Future

Springer

Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings

Quan Zhang · Hong Yang
Editors

Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings

Rasch and the Future

Springer

Editors

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

Hong Yang
College of Foreign Studies
Jiaxing University
Jiaxing, Zhejiang
China

ISBN 978-3-662-47489-1 ISBN 978-3-662-47490-7 (eBook)
DOI 10.1007/978-3-662-47490-7

Library of Congress Control Number: 2015942815

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media
(www.springer.com)

Foreword

Welcome Message from Host University

Prof. Quan Zhang Ph.D, Jiaxing University, China
Secretary of PROMS

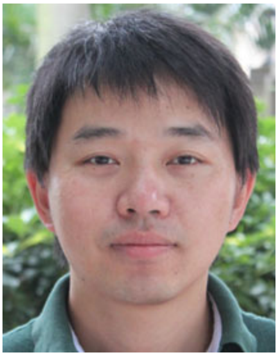
Welcome Message from Chair of PROMS

Prof. Robert F. Cavanagh, President of PROMS
Professor of Well-being Metrics, School of Education
Curtin University, Australia

Pre-Conference Workshops

Workshop I: Rasch Measurement Using WINSTEPS

Prof. Trevor Bond (Keynote Speaker)
The immediate Past President of PROMS
James Cook University, Queensland, Australia



Dr. Zi Yan
Vice President of PROMS
Hong Kong Institute of Education

Workshop II: Introduction to and Demonstration of Using TAM Software for IRT Analysis

Prof. Margaret Wu Ph.D.
Statistician and psychometrician
Victoria University in Melbourne, Australia

Workshop III: Invariant Measurement with Raters and Rating Scales

Prof. Jackson Stenner
Chairman, CEO, and Co-Founder of MetaMetrics Inc. USA

Workshop IV: Introduction to Lexiles and Developing Construct Models

George Engelhard, Jr., Ph.D. The University of Georgia, USA
And his teaching assistant: Shanna N. Ricketts, Emory University, USA

Keynote Speakers

An Argument Approach to Test Fairness: The Case of Multiple-form Equating in the College English Test

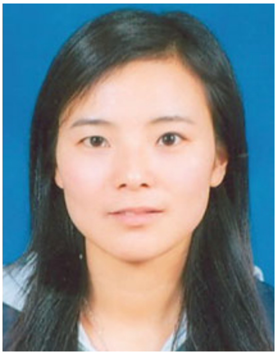
Prof. Yan Jin
Shanghai Jiao Tong University

Dr. Eric Wu
UCLA, USA

DELTA: A System for Diagnosing and Tracking English Competence in Chinese University Students

Alan Urmston
Hong Kong Poly U

Michelle Raquel
Hong Kong Poly U



Gwendoline Guan
Hong Kong Poly U

Foreword

Welcome Message from Host University

I am delighted to announce that the PROMS Board has decided the PROMS2014 symposium will be held in Guangzhou, China from August 4–6 2014 with pre-conference workshops scheduled for August 2–3 2014 and post-conference self-arranged events scheduled for August 7, 2014.

Over the past years, PROMS has been successfully hosted in many Pacific Rim countries and regions for the purpose of promoting research and contributing to the development of the Rasch Model. Following the inaugural PROMS symposium in 2012, held in Jiaxing, Zhejiang Province, China Mainland and the PROMS2013 symposium, Kaohsiung, Taiwan, we are now opening our arms to welcome all PROMS counterparts to Guangzhou, China.

The ideas and concepts regarding the Rasch Model were first introduced into China in the 1980s by Prof. Gui Shichun, my Ph.D. supervisor. Later, it was Prof. Gui who first utilized the Rasch Model in the form of the 10-year long (1990–1999) Equating Project for Matriculation English Test (MET). The results of Prof. Gui's MET received praise and adulation in China, but the following years did not witness significant Rasch-based application and research. With this in mind, we can see the PROMS2014 symposium meets the important need of providing an excellent introduction to the Rasch model and its application.

The PROMS2014 symposium will feature a slightly different schedule from past symposiums. First, the PROMS2014 symposium will have a longer time frame, three days instead of two, for the main conference. To quote Prof. Robert F. Cavanagh, "a three-day program reduces the number of papers presented concurrently. This increases attendance, sharing of work, learning from others, and giving more time for posters, symposia, and fireside chats". A longer program enables more time for less formal activities. Second, bilingual experts will be invited to serve as interpreters and/or coordinators at pre-conference workshops to meet the needs of beginners and students, ensuring everyone gets the same keywords and commands on the screen.

Anyone seriously interested in research and development in the field of psychometrics or measurement will find such an international symposium and related workshops to be an excellent source of information about the application of the Rasch Model. In particular, I should mention the PROMS2014 symposium in Guangzhou would be of great benefit to postgraduate students from developing countries and researchers who seek to use the Rasch Model in their research activities. Academically, the PROMS2014 symposium tends to be more international. Just as Professor Mok said, "Instead of being confined within the Pacific-Rim region, participation of scholars from Europe will make PROMS more 'International'". Furthermore, something worth mentioning is that PROMS2012 proceeding has been officially published by Springer and PROMS2014 will have all the presented papers recommended to Springer, Educational Psychology and others to publish in the form of journal.

Finally, I should say, the city that the PROMS2014 symposium has chosen is unique not for its cosmopolitan size but for its local cuisine. In China it is common to hear people say: "Eating in China, Guangzhou is Number One!"

We look forward to meeting you in Guangzhou!

Prof. Quan Zhang Ph.D.
Senior visiting scholar to ETS
Senior research scholar at UCLA
Institute of Language Testing, University of Jiaxing, Zhejiang Province, China
Ph.D. Supervisor, City University of Macau, SAR, China

A Welcome Message from Chair of PROMS

It is my pleasure to extend a warm welcome to the 2014 Symposium of the Pacific Rim Objective Measurement Society (PROMS). This symposium will be held in Guangzhou, China. Also, I need to sincerely thank Prof. Quan Zhang (Ph.D.), the Dean of the College of Foreign Studies and Director of the Institute of Language Testing at the University of Jiaxing for his initiative and work as the Convenor of PROMS2014.

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 to all participants;

The venue, Guangzhou (Guandong/Canton), is a prosperous metropolis full of vigour located along the south coastline of China, just across the border from Hong Kong, an ideal venue for PROMS business and social activities;

PROMS2014 will be the tenth symposium and follows highly successful meetings in Kuala Lumpur, Hong Kong, Taiwan, Tokyo, Hong Kong, Kuala Lumpur, Singapore, Jiaxing China and Taiwan.

PROMS is a multi-cultural event and typically includes presentations in the native language of the host country.

The core business of PROMS is the application of the Rasch Model in a diverse range of fields including Business, Counselling, Economics, Education, Health Care, Language, Measurement, Psychology, Quality Assurance, Statistics and Strategic Planning. The Society is supported by senior academics, researchers and scientists from Asia, Australia, North America and Europe. All share a vision of sustainable development that is informed by meaningful data and meaningful interpretations of these data. They have a common concern about the inadequacies inherent in many decades of Western Human Science measurement and the need to avoid replication of these shortcomings.

The Symposium program will be 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.

Please join us to celebrate ten years of PROMS. We promise intellectual stimulation; highly enjoyable social activities and an engaging experience for participants and their families.

Professor Robert F. Cavanagh (Ph.D.)
Curtin University; Western Australia
Chair of the PROMS Board of Management

Acknowledgments

The following individuals and sponsors helped make the Pacific Rim Objective Measurement Symposium (PROMS), held in Panyu, Guangzhou, China (August 2–6, 2014), a great success. Their contributions in maintaining the quality of both the paper presentation review and the organization of the international academic conference are greatly appreciated.

Prof. Robert F. Cavanagh (Chair), Curtin University, Australia
Prof. Trevor Bond (Previous Chair), James Cook University, Australia
Prof. Zi Yan (Vice Chair), The Hong Kong Institute of Education
Prof. Quan Zhang (Secretary), University of Jiaxing, China
Prof. Chew, Lee Chin, National Institute of Education, Singapore
Prof. George Engelhard, Emory University, USA
Dr. Iris Lee, National Institute of Education, Singapore
Prof. Ong Kim Lee, National Institute of Education, Singapore
Prof. Magdalena Mo Ching Mok, The Hong Kong Institute of Education, Hong Kong
Prof. Jackson Stenner, MetaMetrics, Inc., USA
Prof. Wen Chung Wang, The Hong Kong Institute of Education, Hong Kong
Dr. Mohd Zali Mohd Nor, Malaysia

Sponsors
Institute of Language Testing, Jiaxing University, China
Foreign Language Teaching and Research Press, Beijing, China
City Training Institute (CTI) Guangzhou, China
MetaMetrics, USA
Springer, Germany

Conference Organizers

Pacific Rim Objective Measurement Symposium 2014 Local Committee
The Local Committee, headed by Prof. Quan Zhang, Ph.D., comprises a team of colleagues including assistant/associate professors and student volunteers who attended to matters such as conference dissemination and promotion, implementation and organizational details, and conference budget.

Conference Chair
Prof. Quan Zhang, Ph.D.
Professor, Faculty of Foreign Studies, University of Jiaxing, China

Organizing Committee

Co-Chairs
Prof. Hong Yang
Associate professor, Faculty of Foreign Studies, University of Jiaxing, China
Prof. Xiaoxi Hu
Professor, Liaison Office, Southern Medical University, Guangzhou, China
Prof. Shuangtian Qian
Professor, City Training Institute (CTI) Guangzhou, China

Members
Dan-wei Chai
City Training Institute (CTI) Guangzhou, China
Xiao-xi Hu
Southern Medical University, Guangzhou, China
Dong Wang
Southern Medical University, Guangzhou, China
Pei-sha Wu
City Training Institute (CTI) Guangzhou, China
Yang Shao
City University of Macau, Macau SAR
Chao-ran Yang
City Training Institute (CTI) Guangzhou, China
Qing-hong Zhang
City Training Institute (CTI) Guangzhou, China
Rou-xin Zhang
University of Jiaxing, China

Mingzhu Miao
Institute of Language Testing, University of Jiaxing, China
Wei chun Zhang
City Training Institute (CTI) Guangzhou, China
Xiao-feng Zhang
City Training Institute (CTI) Guangzhou, China

And all the student volunteers of City Training Institute (CTI) Guangzhou, China.

Pre-Conference Workshops

Pacific Rim Objective Measurement Symposium (PROMS2014) was held in Panyu, Guangzhou, China Mainland from August 5–7, 2014 with four pre-conference workshops listed as follows:

Workshop I: Introduction to Rasch Measurement Using WINSTEPS
3.3.1. Prof. Trevor Bond and Dr. Zi Yan (August 2–3, 2014) who gave a full introduction to Rasch Measurement Using WINSTEPS. Previous knowledge of Rasch model was not required. The two-day workshop covered an introduction to Rasch model, the background knowledge and the basic ideas regarding Rasch measurement including computing ability estimates and item statistics, plotting item characteristic curves, estimating population characteristics and the like. Questions and answers and follow-up discussions were conducted in English with Chinese interpretation.

Workshop II: Introduction to Rasch Measurement Using WINSTEPS
3.3.2. Prof. Margaret Wu (August 2–3, 2014) who gave similar yet more detailed introduction to and demonstration of using TAM (Test Analysis Modules) software for IRT analysis. TAM (Kiefer, Robitzsch and Wu, 2012) is an IRT software program written in R. It is free for download. TAM can fit one-parameter, two parameter and multi-dimensional IRT models and can be used for dichotomous and partial credit item responses. The workshop was run in both English and Chinese to ensure all the participants of non-English majors got the ideas.

Workshop III: Introduction to Rasch Measurement Using WINSTEPS
3.3.3. Prof. Jackson Stenner (August 2, 2014), Chairman, CEO and Co-founder of MetaMetrics Inc, president of the Board of Directors of Institute of Objective Measurement, a board member for the National Institute for statistical sciences and a past board member for Duke Children's Hospital and the North Carolina Electronics and Information technologies Association, USA who gave an introduction to Lexiles and developing construct models.

Workshop IV: Introduction to Rasch Measurement Using WINSTEPS
3.3.4. Prof. George Engelhard (August 3, 2014) who gave a full introduction to Invariant Measurement with Raters and Rating Scales. The use of rating scales by

raters is a popular approach for collecting human judgments in numerous situations. 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 with focuses 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 and also provided an introduction to the Many Facet Model, and its use in the development of psychometrically sound performance assessments. Illustrated were examples based on Advanced Placement English Literature and Composition assessments, as well as other large-scale writing assessments. The Facets computer program (Linacre 2007) was used throughout the workshop to illustrate the principles of invariant measurement with raters and rating scales.

Contents

Chapter 1
Be Wary of What's Coming from the West: Globalisation and the Implications for Measurement in the Human Sciences

Robert F. Cavanagh and William Fisher Jr.

Abstract This work draws together material from three bodies of knowledge with a demonstrable yet rarely explored synergy. These are conceptions of well-being and public good, features of globalisation and neo-liberalism, and the principles and practises of contemporary measurement, the application of the Rasch Model. The stimulus for the analysis was the function of the Pacific Rim Objective Measurement Society (PROMS). According to the PROMS Constitution, the mission of PROMS is to contribute to individual, community, and societal well-being and the public good in the Pacific region. This will be realised by application of Rasch Model measurement to understand the progress and impact of developments in areas such as communication technologies, educational reform, health provision, and welfare delivery. However, this understanding also needs to take into account the influence on these developments of the globalisation agendas of governments and multinational corporations. This paper maps out some ways for Rasch measurement to contribute to the improvement of people's lives and to maximise the benefits of globalisation.

Invited address to the 2014 Symposium of the Pacific Rim Objective Measurement Society: Guangzhou, China.

R.F. Cavanagh (✉)
Curtin University, Bentley, WA, Australia
e-mail: R.Cavanagh@exchange.curtin.edu.au

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

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_1

1.1 Introduction

The year 2014 marks the 10th anniversary of the annual Pacific Rim Objective Measurement Symposium, and it is timely to contemplate what the future might hold for the Pacific Rim Objective Measurement Society (PROMS), its members, and the broader community of Human Science metricians. We commence by examining the mission of PROMS and what it hopes to achieve for people living in the Pacific Rim.

We then adopt a much broader perspective and examine the external environment in which PROMS and Human Science measurement exist. An environment characterised by the inexorable global forces of neo-liberalism globalisation, marketisation, decentralisation, accountability, and performativity. One theme of this address is the need for a discerning approach when contemplating adoption of such reforms. We argue there is both certainty and uncertainty. Propagation and implementation are certain, but the benefits for the adoptees are much less certain. Along similar lines, we draw attention to the hegemonic role of measurement in change and development. Agents of change can use measurement to invoke a sense of legitimacy and confidence, but measurement in itself does not necessarily gauge what is critically important for people or improve their lives.

Next, we present a bold distillation of emergent issues framed as five declarations, a manifesto about the conduct of human measurement applying the Rasch Model. Finally, the discussion is extended with three propositions about the future of metrology in the Human Sciences.

1.2 The PROMS Constitution

We begin with the PROMS Constitution endorsed by the 2013 Board of Management meeting in Kaohsiung Taiwan.

> The principal aims of PROMS shall be to:
>
> - Encourage the pursuit of Rasch measurement in the Pacific region and its application in the fields of Business, Counselling, Economics, Education, Health Care, Language, Measurement, Psychology, Quality Assurance, Statistics and Strategic Planning;
> - Advocate measurement practice that contributes to individual, community and societal well-being, and the public good, in the Pacific region; and
> - Raise the visibility and status of Rasch measurement in the Pacific region.
>
> These aims are to be accomplished through:
>
> - Establishing a Pacific region network of researchers and scholars in order to foster cooperative and collaborative relationships;
> - Holding an annual meeting—the *Pacific Rim Objective Measurement Symposium*;
> - Providing workshops and training; and
> - Engaging in such other activities as might facilitate the development of research, policy and practice in Rasch measurement.

PROMS did not have a constitution prior to the 2011 Symposium at the National Institute of Education in Singapore. The 2011 Board Meeting decided a constitution was needed and Board Members commenced writing drafts based on the constitutions of similar organisations including British Educational Research Association, International Association for Cognitive Education and Psychology Constitution, Pacific Early Childhood Educational Research Association, and World Educational Research Association. While the first drafts of the PROMS Constitution drew upon structure and content from research organisations, these were from the field of education and this had potential for discipline-based limitation. To ensure disciplinary inclusivity, the Constitution welcomes application of the Rasch Model in Business, Counselling, Economics, Education, Health Care, Language, Measurement, Psychology, Quality Assurance, Statistics and Strategic Planning. A multidisciplinary approach is strategically important for increasing synergy between researchers and strengthening the capacity of PROMS to have significant and sustainable impacts.

Early versions of the Constitution referred to 'objective measurement' with no mention of the Rasch Model or of Rasch Measurement. In recent times, there has been an increase in the visibility and scholarly profile of Rasch measurement, particularly through international meetings and definitive publications (e.g. Andrich 1988; Bond and Fox 2007; Engelhard 2013; Wilson 2005). Consequently, the Constitution was amended to explicitly refer to Rasch Measurement. Perhaps in 2014, we might like to be more ambitious in showing our understanding and valuing of Rasch's contribution to measurement by its recognition as a theory of measurement (see Andrich 2011; Engelhard 2013).

The Constitution refers to well-being and the public good. What is well-being? Seligman (2011) identifies elements of well-being theory. Positive emotion (the pleasant life), engagement (about flow—concentration, interest, and enjoyment), meaning (belonging to and serving something that you believe is bigger than the self), accomplishment, and positive relationships. Huppert and Johnson (2010, p. 264) define subjective well-being as "the combination of feeling good and functioning well". Feeling good comprises positive emotions, such as happiness, contentment, interest, and affection. Functioning well includes a sense of autonomy or self-determination, competence and self-efficacy, resilience, and positive relationships. What is the public good? Philosophical schools of thought provide one insight. Here are some philosophies and in some cases, philosophers. These were identified in the positive psychology literature, and the groupings are in approximate historical order. The Eastern philosophies of Confucianism, Taoism, Buddhism, and Hinduism emphasise the need for compassion and harmony. The ancient Western philosophers, Socrates, Plato and Aristotle, analysed human existence with attention to examination of self and experiences, virtues and human strengths, and naturalistically grounded moral and intellectual virtue theories. Early modern philosophy concerned humanism. Locke referred to self-determinism, free will, and willfulness or volition in human thinking. Similarly, Descartes and Voltaire wrote about optimism. The recent modern and unmodern philosophers, Kant, Dewey and Popper, highlighted emancipation and responsibility for one's actions. Finally, the approaches applied by more contemporary philosophers include critical theory with expressions of a socially just world without repression of individuals or of groups.

But the PROMS Constitution charges us with advocating measurement practises that contribute to well-being and the public good. Surely, the objectivity, precision, impartiality, trustworthiness, and so on of the measurement are not compatible with a humanist and subjective view of the world and society? And, even if there is some compatibility, how can this view be manifest in measurement? The resolution of these pseudo-paradoxes is possible by examining the history and philosophy of Western science and scientific thinking. The roots of science are in positivism, a function of the modernist era traceable to the Enlightenment period of the seventeenth and eighteenth centuries. Western civilization sought science as a refuge from politics, and it sought reason as a refuge from force. According to positivism, metaphysics is nonsense, there is a universal and a priori scientific method, there is an objective, independent reality we call the world, truth is correspondence to reality, and scientists discover truth as spectators of a world, which is essentially given. This view of science emanating from the development of the physical and natural sciences is difficult to apply in the human and social science fields. It also restricts the theorising and application of measurement in these fields and the kinds of phenomena that can be measured. Achieving the PROMS aim of measurement to enhance the well-being of individuals, communities, and societies requires a non-positivistic view of science and measurement. In particular, drawing upon post-positivist philosophies in which claims to knowledge are not indisputable (Kuhn 1961; Quine 1951), science and society are intertwined (Heelan 1998; Latour 1990), the substance of scientific knowledge is constructed by scientists (Knorr-Cetina 1983; Latour and Woolgar 1979), and the progress of science is a collective movement (Latour 1987; Galison 1997; Scharff 2011). The PROMS Constitution acknowledges the importance of collaboration and networking including the provision of workshops and training.

A complementary approach to conceptualising the application of science and measurement for the human good comes in the notion of living capital metrics (Fisher 2002, 2007). Some forms of capital (e.g. manufactured, liquid, and property) are brought to life because they can be traded due to the availability of transferable representations of their value. Alternatively, other forms of capital (e.g. human, social, and natural) "remain dead, or as yet unborn, tied as they are virtually everywhere to non-transferable representations-scales with values that change depending on local particulars" (Fisher 2007, p. 1092). The local dependency of this capital inhibits trading because the dependency prevents comparison of how it is valued by different groups. If the measure was invariant, there would be a common system of valuing and communicating value independent of group membership. The invariance associated with measurement leads to creation of common 'currencies' with sufficient temporal and contextual stability for the transactions characteristic of a viable economy. In addition, the calibration of measures across different industries (e.g. Communications, Education, Health, and Welfare) enables the growth of new industries transcending the boundaries of traditional mission statements and operational plans. In Germany, to take an example from the energy sector, small household engines generating heat and electricity are electronically connected and controlled by a centralised processor that regulates energy flow in and out of 1500 houses. The efficient and environmentally sound exchange of energy is determined by algorithms processing data continuously collected throughout the entire system. Multiple parameters are concurrently measured using a common metric and language for communication between the components of the network, a metrological network. For the construction of human science metrological networks, it is the social processes and attributes constituting human capital that need to be understood and assigned value. This situates the measurement process within the broader context of ethical human research, specifically the consequences and limitations of measuring human qualities. Indeed, Fisher (2002, p. 854) notes, "Rasch measurement in and of itself is insufficient for bringing about the fulfilment of the human sciences' potentials for improving peoples' lives". Realisation of the PROMS mission requires more than the application of particular measurement techniques in a controlled environment based on a predetermined scientific design. The adoption of any new technology requires organisations to adapt to new flows and uses of information (Hutchins 2012). Practical use of calibrated instruments in everyday classroom, clinical, social service, and human resource applications requires the coordination of a wide range of other actors' roles, roles not typically included when the task of measurement is addressed, such as accountants, software engineers, administrators, parents, customers, theorists, researchers, and so on. Perhaps PROMS could become a forum for explaining the science informing these technicalities and how it is relevant to the world outside the laboratory as a critical and constructive means of engaging with practical problems.

In summary, we have provided an interpretation of the PROMS Constitution centring on:

- The importance of a multidisciplinary approach;
- Recognition of Rasch Measurement Theory as a scientific theory;
- A post-positivist view of Human Science; and
- Human Science metrological networks.

In the next section, we outline the forces influencing global development, a sketch of the complex environment in which PROMS and Human Science measurement are situated.

1.3 The Globalised Landscape

Five distinguishing characteristics of the modern global environment are presented, neo-liberalism, globalisation, marketisation, decentralisation, and accountability and performativity. Note that these characteristics of the global environment are modern in the philosophical sense of the term and so are largely founded in a positivist perspective. The policy and organisational changes that might be expected to follow from a mindful and ethically oriented implementation of Rasch-calibrated measures are, then, not included in the following description.

Classical liberalism promotes the ideals of personal freedom and "possessive individualism," and it opposes collectivism (Robertson 2007, p. 13). Its key features are individual autonomy, freedom from dependence on others, and the construction of a political society to protect individuals' property (Robertson 2007). In contrast, neo-liberalism is the belief that the market should be the organising principle for all political, social, and economic decisions (Giroux 2005). Neo-liberalism encourages government intervention to nurture and preserve the market model (Hill 2003; Robertson 2007; Thorsen and Lie 2007). The neo-liberal state functions to create and develop an "institutional framework" that ensures free trade, free markets, and strong private property rights (Harvey 2005, p. 2). The impact of neoliberal reform on education has been particularly evident in Australia where it continues to infiltrate schools (Clarke 2012; Davies and Bansel 2007).

Capitalist agencies such as the World Bank and the International Monetary Fund have promoted a model of neo-liberal globalisation that incorporates mass educational reform (Davies and Bansel 2007; Torres 2009). Globalisation is the "rapid acceleration of cross-border movements of capital, goods, labour, services and information" (Green 2003, p. 7). Hursh (2005) suggests that globalisation and neo-liberalism coexist due to the free market notion that economies should be open to free trade and competition to increase efficiency. According to the "neo-liberal globalisation thesis", nation states must liberalize public services to promote global capitalism (Beckmann and Cooper 2004, para. 1). This has resulted in reform in many nations promoting the privatisation of public education (Astiz et al. 2002). In particular, decentralisation has become the standard practice for school governance, an inevitable consequence of both economic and institutional globalisation (Astiz et al. 2002, p. 70).

Marketisation involves the "intensified injection of market principles such as deregulation, competition and stratification into public schools" (Bartlett et al. 2002, p. 1). The marketisation of public education was embraced during the 1980s by conservative administrations in the UK and the USA (Ball and Youdell 2008). This established "quasi-markets" as bureaucratic controls were reduced without a clear price mechanism (Ball and Youdell 2008, p. 18). Through marketisation, policymakers seek to produce competition between schools in expectation of raising standards and promoting efficiency (Ball and Youdell 2008, p. 18). Marketisation also occurs through outsourcing aspects of schooling, such as school functions or administration duties, for-profit organisations, and the proliferation of school–business partnerships (Bartlett et al. 2002). Ball and Youdell (2008) suggest that the education markets created by neo-liberal policy are not authentic "free markets," as they tend to be subjected to considerable "regulation, direction and involvement by the state" (p. 19).

Proponents of decentralisation claim that in addition to the efficiency benefits bestowed by a market model, local management improves student achievement (Cobbold 2012; Loeb et al. 2011). Cobbold (2012), however, refutes this claim in a review of decentralised education systems around the world. He suggests decentralised systems of schooling have not performed better than traditional public schools in 20 years of reform around the world (Cobbold 2012). Other studies, including an Organization for Economic Cooperation investigation of markets in education, have also found that there is no clear evidence that decentralised models of schooling raise student achievement in comparison to traditional state-provided schooling (Ball and Youdell 2008; Waslander et al. 2010).

According to Ball and Youdell (2008), accountability and performance management mechanisms from business are being transferred to the public sector. As governments expand the market economy into education, efficient business practices require the establishment of cost-effective accountability mechanisms (Butland 2008, p. 5). These mechanisms are implemented to provide evidence of entrepreneurial efficiency and ensure that educational processes are transparent and accountable (Apple 2004; Ball and Youdell 2008). Accountability measures include benchmarking schools, publication of school performance, and tying teacher pay to student outcomes (Ball and Youdell 2008). Butland (2008) indicates that an emphasis on "large-scale testing and standards" has provided "accountability data" as part of a global phenomenon that applies market theory to education (p. 4). Ball (2003) suggests that the accountability epidemic can be aligned with the "technology of performativity" (p. 216). This is the process whereby the "performances" of individuals or organisations serve as measures of productivity or output (Ball 2003, p. 216). Performativity "employs judgements, comparisons and displays as means of incentive, control, attrition and change based on rewards and sanctions" (Ball 2003, p. 216). Measures of productivity such as "analysis of students performance data" including national standardised test data are used to determine teacher performance (Australian Labor 2012). As part of a national accountability framework, this "accountability data" is publicly available to the community in the form of standardised test results (Redden and Low 2012).

Next, we survey the intersection of this global turn with contemporary Human Science measurement. Five declarations about the conduct of Human Science measurement are written to elucidate defensible principles and practices.

1.4 Towards a Manifesto of Human Science Measurement

A design-oriented approach to Rasch Measurement begins from an a priori theoretical model of the construct of interest that is specified before instrument construction, data collection, or data analysis (Bunderson and Newby 2009; Stenner et al. 2013; Wilson 2005). Typically, the construct model will be based on previous research, theory, or experience and describe components and relations constituting a unidimensional phenomenon. Despite the advantages that accrue from intentionally designing instruments likely to produce data fitting a Rasch model in accord with a construct theory (Fisher 2006), it is common practice for model specifications and construct maps to be retrospectively determined from pre-existing instruments, data, and analyses. Qualities of persons and groups are explored, clarified, and operationally defined. Particular behaviours, abilities, dispositions, attitudes, beliefs, values, expectations, or circumstances can be selected for incorporation in the model. However, in many instances, the human qualities of interest are ideologically situated within social and economic systems and subject to cultural and subcultural influences. Conceptions of well-being and human development become contestable leading to highly equivocal construct models with subsequent problems for instrumentation. Of equal importance are the moral and ethical imperatives in the conduct of Human Science research, particularly the development of the Aristotelian human good and improving the human condition.

Our first declaration addresses these issues. *The starting point for Human Science research is the understanding of culturally sensitive epistemologies through a concern for the welfare and rights of persons.*

A Rasch Model construct map requires identification of increasing and decreasing amounts of the object of measurement (e.g. the ability, attitudes, and satisfaction of persons). The difficulties instrument items present to subjects need to follow a similar rule. In addition to revealing the qualitative differences between persons, these differences are ordered. The construct map displays both the differing amounts and their order. In addition to identifying types of qualities, the hypothesised degree of manifestation of a quality is provided. The construction of measures with these properties is pre-empted by an understanding of growth, phases of development, or innovation progress. For example, the implementation stages in the privatisation of a government organisation, in the decentralisation of state-owned financial institutions, in the outsourcing of public transport, or in the rationalisation of electricity supplies.

Our second declaration is about meaningful measurement of the progress of reforms for individuals and communities. *Human Science measurement requires construction of linear theoretical models to map out transformations in developments impacting on the quality of persons' lives.*

The globalisation of societies and industries presses for comparisons of human activity on a global scale. The Programme for International Student Assessment collects data on student achievement in mathematics, science, and reading from 65 nations and territories. The instruments used for cross-national contrasts need to function independent of potentially confounding situational variables to display invariance. Instruments or measures are said to be invariant, in the ideal sense of a pragmatic heuristic fiction, when many instruments produce a similar reading for the same amount of a quantity, or one instrument produces the same reading for the same amount of a quantity but in different locations or situations. In educational testing, the ordering of students is hypothesized to not depend on which items were administered (invariant person measures). Or, when one form of a test is administered to students in different places or at different times, students with similar ability should attain similar measures. The ordering of the difficulties of the items present to students should not depend, within the relevant range of uncertainty, on which students wrote the test (invariant item difficulties). The notion of invariance is also applicable in the measurement of health, prosperity, happiness, and other human attributes (Andrich 2011; Bezruzcko 2005; Bond and Fox 2007; among many others). It is also a requirement of the Rasch Model acknowledged to be unrealistic and untrue, but useful (Box 1979; Rasch 1960), as has long been recognized and accepted relative to the status of laws and models in the natural sciences (Butterfield 1957; Cartwright 1983; Holton 1988).

Our third declaration is about globalisation and the performance of international measures. *The instruments and data required in the management of global programs should be proven invariant across different nations and societies.*

Implicit in the previous sections is the existence of Human Science measures, the presence of numbers, and units of measurement to provide meaningful representations of quantities of human qualities. The meaning conveyed in a measure derives from a common and agreed-upon understanding of the unit of measurement. Units in the Physical Sciences have developed from extensive attention to theory and involved many scientists over long periods. While the underlying metrological systems may not be visible to end users, quantities using these units are universally comprehendible. The availability of units of measurement is not yet a matter of widespread concern in the Human Sciences. An exception to the lack of concern for invariant units is Rasch measurement. In Rasch measurement, the unit of measurement, the logit, can be estimated from observed scores or theoretical calibrations to mark person measures and item difficulty locations on an interval scale. A person's measure is estimated from the probability of that person being able to complete certain tasks (the items). The meaning of the measure derives from the nature of these tasks and the order of their difficulty. The logit measure directs the measurer to a certain point on the scale which indicates the tasks the subject is expected to be able to complete successfully. The meaning of a person measure expressed in logits is elucidated by reference to a map of task difficulties, known variously as a construct map, a Wright map, or a variable map (Masters et al. 1994; Wilson 2005).

Our fourth declaration is about the meaning of measures. *Units of measurement defined by maps of task difficulty and theoretical explanations of that difficulty are required for meaningful quantification in the Human Sciences.*

Meaningful measurement cannot ensure but aids in making the quantity measured visible and tangible to end users. The substance of the measure is the explicit meaning of the unit of measurement, the manifestation of more and less. Provided there is sufficient substantive similarity, units can be incorporated into systems and task descriptors marked on common scales (not to be confused with the traditional calibration of instruments using correlational methods and requiring complete data). These systems make possible networks of Human Science metrics with global exchange of data between scientists and across disciplines. The networks provide Fisher's (2002, 2007) transferable representations and mechanisms for liberating human capital, in the sense of making it meaningfully actionable at local levels in global terms.

Our fifth declaration is about metrological networks in the Human Sciences. *The contribution of Human Science measurement to economic and social development requires collaborative construction, maintenance and growth of metrological networks.*

Metrologists have lately sought to bring psychological and social measurement into a common frame of reference with physical measurement (Joint Committee for

Guides in Metrology 2008; Mari and Wilson 2013; Pendrill and Fisher 2013). Following Latour's (2005) Actor Network Theory (ANT), furthering this effort at collaborative construction, maintenance, and growth of metrological networks will require making three significant moves, moves that are conceptually and organisationally innovative and so, challenging, as there are no, or few, historical precedents to adopt and adapt as models. If Rasch measurement researchers would adopt these moves as components of their methodological orientation, perhaps new developments in the direction of metrological traceability for psychological and social measures would ensue in the form of a practical program complementing existing ANT-based critiques of education (Fenwick 2010).

The first move is localizing the global (Latour 2005). In Rasch measurement terms, this means stating the construct model by explicitly parameterizing the potential universal, designing an instrument that embodies the construct theory, testing specific hypotheses as to the local existence of the universal, gathering data using the instrument, performing experimental evaluations of scoring hypotheses, and assessing the extent to which the local expression of the construct is specific to the particular time and place and the extent it might be participating in a global universal. Graphical construct maps that make both the measures and the exceptions that prove (in the sense of testing) the rule visible and interpretable to end users and researchers alike are key tools (Chien et al. 2009; Masters et al. 1994; Wilson 2005). Rasch measurement exceeds the plan as articulated by Latour, as Latour is focused on the two-dimensional mapping of relationships in an information network and completely ignores the electronic informational content varying along a single dimension that is what is transmitted through that network. Of course, tracing relationships based on the communication and performance of test scores instead of Rasch measures would lead to very different networks and consequences. It will, however, be important in projecting new possibilities for working out an extended net of relationships to focus on the connected sites and not to worry about their relative sizes or ranking them into micro and macro.

The second move is redistributing the local (Latour 2005). The question is, how mobile is the so far singularly located potential, inchoate, or stochastic universal? Mobility implies both conceptual redistribution across particular collections of local questions asked and redistribution across persons participating in the conversation. Does the potential universal maintain its integrity across samples of questions asked and persons answering? Can a theory of the construct inform the formulation and posing of questions as well as data does? How stable is the assemblage of questions? How much can it be moved without deforming the invariant construct profile, what Latour (1987) calls the "immutable mobile"? Can a constant connection be maintained between distant locations? What does it take in the way of distorted question content, administration, or sampling to break the connection? How diagnostically useful are individual inconsistencies that depart from the modelled expectations? Documented answers to these questions in published Rasch research would be an immense aid to the cultivation of a new metrological culture aware of its potentials and challenges.

The third move is connecting sites (Latour 2005). Given the repeated and reproducible pattern of an invariant stochastic universal, what needs to be done to connect the nodes of the redistributed local into a "worknet" capable of sustaining the common unity embodied in the communications of communities of research and practice? What applications capitalizing on the invariance suggest themselves or can be imagined? How can measures produced instantaneously at the point of use in the classroom, clinic, office, or home be put to immediate use in diagnosing idiosyncratic problems or in determining what comes next up the scaffold? How can statistical summaries of previously made measures be put to work as contextual contrasts informing the interpretation of new measures? What are the implications for tracking growth, development, change? What opportunities for continuous quality improvement and teacher collaboration emerge when measures across all fifth grade mathematics classrooms are comparable? Again, taking advantage of existing electronic networks seems to be the order of the day when following through with the momentum of a Rasch measurement application.

1.5 Conclusion

This paper was organised around three constructs and the pragmatic and theoretical dynamics between these. First, the notions of well-being and the public good; these can be traced back to ancient times, are identifiable in many religions and philosophies, and are the subject of contemporary theorising and research. Second, measurement; this is integral to science and also to technology, industry, and commerce. Measurement was presented as a social process with the potential for a positive impact deriving from the availability of common and invariant metrics. These metrics can be built into networks in which data are shared and new representations of nature and human relations emerge. Third, the context for human and industrial development was conceptualised as five aspects of globalisation.

The three-way dynamic was explored by identifying the imperatives and implications of the construction of Human Science measures, specifically, measures applying the Rasch Model and Rasch Measurement Theory. These were considered in conjunction with the well-being of persons and the public good and with globalisation. Scrutiny of the confluences produced the following five declarations about Human Science research and Rasch Theory measurement:

1. The starting point for Human Science research is the understanding of culturally sensitive epistemologies through a concern for the welfare and rights of persons.
2. Human Science measurement requires construction of linear theoretical models to map out transformations in developments impacting on the quality of persons' lives.
3. The instruments and data required in the management of global programs should be proven invariant across different nations and societies.

4. Units of measurement defined by maps of task difficulty and theoretical explanations of that difficulty are required for meaningful quantification in the Human Sciences.
5. The contribution of Human Science measurement to economic and social development requires collaborative construction, maintenance, and growth of metrological networks.

We concluded by drawing on Actor Network Theory (Latour 2005) to propose three moves necessary for the creation of comprehensive metrological networks in the Human Sciences. First was 'localizing the global' by developing a global and universally understood representation of a phenomenon, operationalising this conception in ways amenable for verification locally and then evaluating the local and global constructs using local data. Second was 'redistributing the local' by ascertaining the retention of the core substance of the universal representation when the universal is transported to different locations and ascertaining how much localised interpretation can be tolerated before irreversible distortion of the universal. Third was the practicality of 'connecting sites' or localised network nodes, sharing knowledge of the invariant universal and nurturing the growth of communities bonded by common measurement approaches and practices.

We have argued for recognition of a symbiotic relation between well-being and the public good, globalisation, and Rasch Measurement Theory. We present what we believe to be a promising and positive outlook for critical and constructive progress in measurement research and practice.

Acknowledgments The authors are extremely grateful to Carly Keay, a post-graduate student at Curtin University. The treatment of globalisation in this paper was from her research and writing.

References

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Reviews Pharmacoeconomics Research, 11*(5), 571–585.

Apple, M. W. (2004). Creating difference: Neo-liberalism, neo-conservatism and the politics of educational reform. *Educational Policy, 18*(1), 12–44. doi:10.1177/0895904803260022.

Astiz, M. F., Wiseman, A. W., & Baker, D. P. (2002). Slouching towards decentralization: Consequences of globalization for curricular control in national education systems. *Comparative Education Review, 46*(1), 66–88.

Australian Labor (2012). *Reward payments for great teachers*. Retrieved from http://www.alp.org.au/agenda/education—training/performance-pay/.

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy, 18*(2), 215–228. doi:10.1080/0268093022000043065.

Ball, S. J., & Youdell, D. (2008). *Hidden privatization in public education*. Education International. Retrieved from http://scholar.googleusercontent.com/scholar?q=cache:kh0QIkVqgrEJ:scholar.google.com/+stephen+ball+neoliberal+education&hl=en&as_sdt=0,5&as_vis=1.

Bartlett, L., Frederick, M., Gulbrandsen, T., & Murillo, E. (2002). The Marketization of education: public schools for private ends. *Anthropology & Education Quarterly, 33*(1), 1–25. doi:10.1525/aeq.2002.33.1.5.

Beckmann, A., & Cooper, C. (2004). Globalisation, the new managerialism and education: Rethinking the purpose of education in Britain. *The Journal for Critical Education Policy Studies*, *2*(2). Retrieved from http://www.jceps.com/?pageID=article&articleID=31.

Bezruczko, N. (Ed.). (2005). *Rasch measurement in health sciences*. Maple Grove, MN: JAM Press.

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

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–235). New York: Academic Press Inc.

Bunderson, C. V., & Newby, V. A. (2009). The relationships among design experiments, invariant measurement scales, and domain theories. *Journal of Applied Measurement, 10*(2), 117–137.

Butland, D. (2008). *Testing times: Global trends in marketisation of public education through accountability testing*. NSW Teachers Federation. Retrieved from http://www.pandc.org.au/files/uploads/DButlandPaper.pdf.

Butterfield, H. (1957). *The origins of modern science (revised edition)*. New York: The Free Press.

Cartwright, N. (1983). *How the laws of physics lie*. New York: Oxford University Press.

Chien, T. -W., Wang, W. -C., Wang, H. -Y., & Lin, H. -J. (2009). Online assessment of patients' views on hospital performances using Rasch model's KIDMAP diagram. *BMC Health Services Research, 9*, 135 [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727503/].

Clarke, M. (2012). Talkin' 'bout a revolution: the social, political, and fantasmatic logics of education policy. *Journal of Education Policy, 27*(2), 173–191. doi:10.1080/02680939.2011.623244.

Cobbold T. (2012). *School autonomy is not the success claimed. Save Our Schools.* Retrieved from www.saveourschools.com.au/file_download/100.

Davies, B., & Bansel, P. (2007). Neoliberalism and education. *International Journal of Qualitative Studies in Education, 20*(3), 247–259. doi:10.1080/09518390701281751.

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

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. (2002). The "Mystery of Capital" and of the human sciences. *Rasch Measurement Transactions, 15*(4), 854–857.

Fisher, W. P., Jr. (2006). Survey design recommendations [expanded from Fisher, W. P. Jr. (2000) *Popular measurement, 3*(1), pp. 58–59]. *Rasch Measurement Transactions, 20*(3), 1072-1074.

Fisher, W. P., Jr. (2007). Living capital metrics. *Rasch Measurement Transactions, 21*(1), 1092–1109.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.

Giroux, H. A. (2005). Cultural studies in dark times: Public pedagogy and the challenge of neoliberalism. *Fast Capitalism*, *1*(2). Retrieved from http://www.fastcapitalism.com/.

Green, A. (2003). Education, globalisation and the role of comparative research. *London Review of Education, 1*(2), 84–97. doi:10.1080/14748460306686.

Harvey, D. (2005). *A brief history of neoliberalism*. Oxford; New York: Oxford University Press.

Heelan, P. A. (1998). The scope of hermeneutics in natural science. *Studies in History and Philosophy of Science, 29*(2), 273–298.

Hill, D. (2003). Global neo-liberalism, the deformation of education and resistance. *Journal for Critical Education Policy Studies*, *1*(1). Retrieved from http://www.jceps.com/?pageID=article&articleID=7.

Holton, G. (1988). *Thematic origins of scientific thought: Kepler to Einstein (Revised ed.)*. Cambridge, Massachusetts: Harvard University Press.

Huppert, F. A., & Johnson, D. M. (2010). A controlled trial of mindfulness training in schools: The importance of practice for an impact on well-being. *The Journal of Positive Psychology: Dedicated to furthering research and promoting good practice, 5*(4), 264–274.

Hursh, D. (2005). Neo-liberalism, markets and accountability: Transforming education and undermining democracy in the United States and England. *Policy Futures in Education*, *3*(1), 3–15. Retrieved from firgoa.usc.es/drupal/files/hursh.pdf.

Hutchins, E. (2012). Concepts in practice as sources of order. *Mind, Culture, and Activity, 19*, 314–323.

Joint Committee for Guides in Metrology (JCGM/WG 2). (2008). *International vocabulary of metrology: Basic and general concepts and associated terms, 3rd ed.* Sevres, France: International Bureau of Weights and Measures–BIPM. http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf.

Knorr-Cetina, K. (1983). Ethnographic study of scientific work: Towards a constructivist interpretation. In Karin Knorr-Cetina & Michael Mulkay (Eds.), *Science observed: Perspectives on the social study of science*. London: Sage.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52*, 161–190.

Latour, B. (1987). Science in action. Cambridge, Mass: Harvard University Press. *Studies in History and Philosophy of Science, 21*(1), 145–171.

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. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills, California: Sage.

Loeb, S., Valant, J., & Kasman, M. (2011). Increasing choice in the market for schools: Recent reforms and their effects on student achievement. *National Tax Journal*, *64*(1), 141–164. Retrieved from cepa.stanford.edu/sites/default/files/A06-Loeb.pdf.

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.

Masters, G. N., Adams, R. J., & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research, 21*(6), 595–610.

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.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*, 20–43.

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.

Redden, G., & Low, R. (2012). My school, education, and cultures of rating and ranking. *Review of Education, Pedagogy, and Cultural Studies, 34*(1–2), 35–48. doi:10.1080/10714413.2012.643737.

Robertson, S. L. (2007). Remaking the world: Neo-liberalism and transformation of education and teachers' labour. *Centre for Globalisation, Education and Societies*. Retrieved from http://www.bris.ac.uk/education/people/academicStaff/edslr/publications/17slr.

Scharff, R. (2011). Displacing epistemology: Being in the midst of technoscientific practice. *Foundations of Science, 16*(2), 227–243.

Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. New York, NY: Free Press.

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 [doi: 10.3389/fpsyg.2013.00536].

Thorsen, D. E., & Lie, A. (2007). *What is Neoliberalism*. Oslo, Norway: Oslo universitesforlaget. Retrieved from http://folk.uio.no/daget/What%20is%20Neo- Liberalism%20FINAL.pdf.

Torres, C. A. (2009). *Education and neoliberal globalization*. New York: Routledge. Retrieved from http://www.scribd.com/doc/82213992/Ebooksclub-org-Education-and-Neoliberal-Globalization-Routledge-Research-in-Education.

Waslander, S., Pater, C., & Weide, M. van der. (2010). *Markets in education: An analytical review of empirical research on market mechanisms in education.* OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/content/workingpaper/5km4pskmkr27-en.

Wilson, M. (2005). *Constructing measures: An item response modelling approach*. New York, NY: Routledge.

Chapter 2
Rasch Model: Status Quo and Prospect in China

Quan Zhang and Tiantian Zhang

Abstract This paper presents an overview regarding the use of Rasch model for research currently across China, listing some significant Rasch-based research work being done or to be done ever since Pacific-Rim Objective Measurement Symposium 2012 held in China. In a word, Rasch-based research work is increasing in China. The status quo is optimistic and prospect tantalizing.

Keywords Rasch model · Translation · PROMS · Conference · Workshop

2.1 Rasch Model: Status Quo in China

Ever since Pacific-Rim Objective Measurement Symposium (PROMS) (Zhang and Yang 2012) was conducted in Jiaxing, China, with the influence, efforts have been taken to undertake the Rasch-based research work in three aspects as follows: research project, translation work, and workshop and conference.

Invited address to the 2014 Symposium of the Pacific Rim Objective Measurement Society: Guangzhou, China.

Q. Zhang (✉)
College of Foreign Studies, University of Jiaxing, Jiaxing, Zhejiang,
People’s Republic of China
e-mail: qzhang141@aliyun.com; proms2012_reg@hotmail.com

Q. Zhang
Institute of Language Testing, University of Jiaxing, No.56 YueXiuNanLu,
Jiaxing 314001, Zhejiang, People’s Republic of China

Q. Zhang
City University of Macau, SAR, Macau, People’s Republic of China

T. Zhang
Faculty of Department of Computer Animation and Engineering, Guangdong
Light Industry Engineer College, Guangzhou, People’s Republic of China
e-mail: gdivf@163.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_2

2.1.1 A Rasch-Based Equating Project with Educational Assessment Australia

This is a joint research work being conducted by both Institute of Language Testing Jiaxing University and Educational Assessment Australia (EAA).[1] The relevant data have been successfully collected, and the report with detailed analysis via UMM and Gitest is to be submitted for presentation at PROMS 2015 in Fukowa, Japan. It is believed that such an equating project would be beneficial for Chinese students who are preparing for their national entrance examination from 2016 onward.

2.1.2 A Rasch-Based Project for Comparison Between College English Test and General English Proficiency Test

Besides the joint project with EAA, a particular Rasch-based project[2] was recently conducted ad hoc for comparison of English listening and reading comprehension between two important English language tests, i.e., General English Proficiency Test (GEPT) in Taiwan and College English Test (CET) in China Mainland (Zhang and Yang 2014; Zhang et al. 2014; Yang and Miao 2014).

2.1.2.1 Test Descriptions

GEPT in Taiwan is a test of English proficiency with five levels currently being administered: elementary, intermediate, high-intermediate, advanced, and quality, of which the high-intermediate level is administered ad hoc to university students of non-English majors.[3]

CET Band 4 is a test of English proficiency for educational purpose designed by Shanghai Jiaotong University according to the requirements of CET and administered to sophomore students of non-English majors only. CET has been administered ever since 1987 across China Mainland and even beyond. The count of CET test takers remains number one in today's world.[4]

[1] This project is financially supported by EAA and was conducted in China in 2013. The details will be jointly presented at the forthcoming PROMS2015, Fukuoka, Japan.

[2] For details, please refer to Chapters 9–10 of the book.

[3] For more details, please visit http://www.gept.org.tw/.

[4] For more details, please visit http://www.cet.edu.cn/.

2.1.2.2 Research Purpose

The motivation to compare these two tests remains for ages. Online search has revealed that so far no significant researches were ever conducted in this regard. And PROMS 2013 held in Kaohsiung sparked the action and made such a research feasible. Also, the time is mature to do such a comparison with focus on both test takers' ability and test item's difficulty in terms of listening and reading comprehension in the Chinese context. Finally, Rasch model turns out to be the most appropriate approach to fulfill the task.

2.1.2.3 Significance and Limitations

The research project, while focusing on the comparison of GEPT and CET, presents the research method via Rasch supported with real data analyses and thus can be concluded in at least two points as follows.

At the first place, the present study is the pioneer one ever conducted in language testing field across Taiwan Strait. Next, probably the most significant parts of the present research are listed as follows: (1) to show to our teachers of English the importance of item analysis and test scoring with the help of Rasch (2) to demonstrate how item analysis and test scoring are actually conducted using GiTest, and (3) to understand the ideas regarding Rasch Model with detailed interpretation. However, two corresponding limitations exist: small sample size and further justification needed to administer the same test items to the students of homogeneous background in Taiwan in the same manner.

2.2 Translation Work

With joint efforts, Rasch work translation is going on well across China Mainland, Hong Kong Macau, and Taiwan. The focus is on Journal of Applied Measurement (JAM) and monograph on Rasch.

2.2.1 JAM Book of Abstract Translation

Led by Profs. Magdalena Mo Ching MOK and Zhang Quan, the series translation of books is going fine. Totally, 7 volumes will be published. So far, Volume II, Constructing Variables, Book of abstract translation of JAM from English into Chinese has been completed and published. Volume I is yet to be published to meet the readers soon. Volume II contains 205 abstracts from JAM germane to Rasch-based research work translated from English to both simple and classic Chinese by 45 highly competent translators who are working in 13 different organizations, universities, or institutes located, respectively, in China Mainland, Hong Kong, Macau, and Taiwan. Each of these abstracts deals with Rasch measures in their research field, covering a variety of issues ranging from education, psychology, management, testing to medicine, and serving in particular as good resources for researchers and students of non-English majors in China Mainland to be able to conduct their own Rasch model analyses as well as understand and critique published Rasch-based research.

2.2.2 Monograph Translation

So far the translation from English into Chinese of the monograph[5] by Prof. Trevor Bond has been completed and is to be published in 2016. It took approximately 3 years. The book contains all the chapters germane to Rasch theory and ideas by eight highly competent/Rasch translators who are working in five different organizations, universities, or institutes located, respectively, in China Mainland and Hong Kong. The translation serves in particular as good resources for researchers and students of non-English majors in China Mainland to be able to learn and self-teach Rasch so as to conduct their own Rasch-based analyses as well as understand and critique published Rasch-based research. Although highly theoretic, the book can be used as a text book for postgraduate students of applied linguistics in China. To quote a few words from the preface written by John "Mike" Linacre to illustrate the point: "The first edition of this remarkable work arrived stealthily. Those of us in the know were aware that it was to be released at the 2001 AERA Annual Meeting in Seattle. When the Exhibitor area opened, I headed for the Lawrence Erlbaum Associates booth and looked for the book. I purchased the very first copy. By the end of the AERA Meeting, "Bond & Fox" had sold out and was on its way to becoming an Erlbaum best seller. And deservedly so. Rigorous measurement has been essential to the advancement of civilization since Babylonian times, and to the advancement of physical science since the Renaissance."

2.3 Workshops and Conferences

Apart from the translation work, more time, funding, and energy are also devoted to hosting international conferences such as PROMS 2012 in Jiaxing, PROMS 2013 in Kaohsiung, and PROMS 2014 in Guangzhou. Each time, experts of Rasch are invited to run pre-conference workshops to offer the practice of Rasch measurement to young teachers, researchers, and students coming from Pacific-rim regions and

[5] Trevor Bond and Christine M. Fox. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences Paperback ISBN-13: 978-0805842524 ISBN-10: 0805842527.

countries and beyond. In what follows is listed a brief introduction to each pre-conference and workshops run by each conference.

1. Pre-conference workshop and PROMS 2012, Jiaxing
   PROMS 2012 (Zhang and Yang 2012) was held in Jiaxing University, Zhejiang Province, China, from August 6–9, 2012, with four pre-conference workshops run by:

   1.1. Professor Robert F. Cavanagh (August 5, 2012) who focuses on Rasch Model for beginners. The workshop attracts more Chinese participants and Rasch beginners from Pacific-rim regions and countries. And the interpretation was provided in Mandarin Chinese. This is the first time that PROMS came into China Mainland ever since it was first held in Malaysia in 2005.
   1.2. Professor Trevor Bond (August 4–5, 2012) who presented with interpretation support in Mandarin Chinese to let more early Mainland researchers to fully understand the idea of Rasch Model used in measurement. Young researchers recommended by the workshop runner also gave some strand keynote. WINSTEPS was demonstrated with examples to illustrate the ideas.
   1.3. James Sick, EdD (August 5, 2012) from International Christian University, Tokyo, Japan, who gave introduction to many facets of Rasch measurement using FACETS. At the workshop, participants took their personal laptops. A time-limited edition of FACETS and example data for use were provided by Mike Linacre. Interpretation was provided in Mandarin Chinese for better understanding of non-English major participants.
   1.4. Dr. Eric Wu (August 5, 2012) from UCLA who gave a full introduction to EQS At the workshop, participants took their personal laptops. B version of EQS and example data for use were demonstrated. His workshop was run in both English and Mandarin Chinese.

2. PROMS 2013 was held in Kaohsiung, National Sun Yat-sen University, Taiwan, from August 3–5, 2013, with three pre-conference workshops run by:

   2.1 Dr. Mark Ronald Wilson (August 2, 2013) from University of California, Berkeley, USA, whose topic is “The BEAR Assessment, Large Scale Assessment, and Explanatory Measurement.” The workshop is divided into three parts. The first part introduces the BEAR Assessment System (BAS) designed to build upon methodological and conceptual advances in assessment. The second part explores implications of the BAS for large-scale assessment. The argument here lies in that the current emphasis on testing in education has led to a, to use Mark’s word, “squeeze” that teachers and their students experience between the need to cover many standards and the educator’s wish to value meaningful instruction and assessment. In the third part, the presenter describes a role for the BAS in explanatory measurement, emphasizing how item response models can be coordinated and broadened to stress their explanatory uses beyond their standards descriptive uses with the ideas exemplified in the context of a reading comprehension test.

2.2 Professor Jack Stenner (August 1, 2013) who ran LEXILE Workshop with research on and application of the LEXILE framework for reading abstract. Professor Stenner provided extensive, interactive training on how the LEXILE framework for reading was built. Participants attended the workshop in group activities with hands-on learning experiences that can impact their day-to-day assessment and psychometric practice.

2.3 Professor Margaret Wu (August 1–2, 2013) who gave detailed introduction to and demonstration of Using Test Analysis Modules (TAM) software for IRT analysis. TAM fits the Rasch model and 2PL models that can estimate unidimensional and multidimensional models with latent regression and facet terms. Both joint maximum likelihood and marginal maximum likelihood estimations can be used. Detailed instructions regarding downloading TAM were provided before the workshop.

3. PROMS 2014 (Zhang and Yang 2014) was held in Panyu, Guangzhou, China Mainland, from August 5–7, 2014, with four pre-conference workshops run by:

3.1 Professor Bond, Trevor and Dr. Yan, Zi (August 2–3, 2014) who gave a full introduction to Rasch Measurement Using WINSTEPS. Previous knowledge of Rasch model is not required. The 2-day workshop covered an introduction to Rasch model, the background knowledge, and the basic ideas regarding Rasch measurement including computing ability estimates and item statistics, plotting item characteristic curves, estimating population characteristics, and so on. Questions and answers and follow-up discussions were conducted in English with Chinese interpretation.

3.2 Professor Margaret Wu (August 2–3, 2014) who gave similar yet more detailed introduction to and demonstration of Using TAM software for IRT analysis. TAM is an IRT software program written in R. It is free for download. TAM can fit one-parameter, two-parameter, and multidimensional IRT models and can be used for dichotomous and partial credit item responses. The workshop was run in both English and Chinese to ensure all the participants of non-English majors got the ideas.

3.3 Professor Stenner, Jackson (August 2, 2014), chairman, CEO, and co-founder of MetaMetrics Inc, president of the Board of Directors of Institute of Objective Measurement, a board member for the National Institute for statistical sciences and a past board member for Duke Children's Hospital and the North Carolina Electronics and Information technologies Association, USA who gave introduction to LEXILES and developing construct models.

3.4. Professor Engelhard, George (August 3, 2014) who gave a full introduction to Invariant Measurement with Raters and Rating Scales. The use of rating scales by raters is a popular approach for collecting human judgments in numerous situations. This workshop utilizes the principles of invariant measurement (Engelhard 2013) combined with lens models from cognitive psychology to examine judgmental processes that arise in rater-mediated assessments, with focuses 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 is 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 extends these ideas to measurement situations that require raters to make judgments regarding performance assessments and also provides an introduction to the Many Facet Model and its use in the development of psychometrically sound performance assessments. Illustrated are examples based on Advanced Placement English Literature and Composition assessments as well as other large-scale writing assessments. The FACETS computer program designed by Linacre in 2007 is used throughout the workshop to illustrate the principles of invariant measurement with raters and rating scales.

4. Furthermore, PROMS 2012 and 2014 Conference Proceeding (Zhang and Yang 2012, 2014) published by the Springer are listed by Conference Proceeding Citation Index (CPCI). In this way, Rasch measurement is further disseminated across China, in the Pacific-rim and in the world as well. The book performance report published by the Springer shows the idea.[6]

   Here subsequently is provided an overview of how PROMS 2012 Proceeding has been performing on the market. Because e-Books have become well established among academic and corporate scientists, the paper-based report concentrates on the electronic version of your publication. Our PROMS Proceeding e-Book is available from Springer Link, which provides readers with access to millions of scientific documents from journals, books, series, protocols, and reference works. All types of publications are interconnected and are fully indexed and searchable to chapter level. As a result PROMS 2012 Proceeding e-Book appears as high up on search engine results' lists as possible and gains higher visibility.

   ISBN 978-3-642-37592-7(e-book) ISBN 978-3-642-37591-0 (print book) Book Performance Report springer.com.

2.4 Rasch Model: Prospect in China

It goes without saying that the increasing Rasch-based research work in China is by no means confined within the outline mentioned earlier. More examples can be observed: in Hong Kong, apart from the PROMS pre-conference workshops, workshop of the similar nature was also arranged on January 15–16, 2015. "A General Class of Latent Structure Models Useful for Applications in Cognitive

[6] http://www.springer.com/alert/urltracking.do?id=L4b93afaMf771e4Sb0b3f0a.

Diagnosis, Scaling, and Clustering" was organized by Assessment Research Centre, The Hong Kong Institute of Education. In Jiaxing, Rasch Model Online Forum has been launched and in Macau, the new PhD program, City University of Macau has started to get involved in application of Rasch for CAT research, a new PhD program approved by the Ministry of Education, China. In a word, the status quo is optimistic and the prospect is tantalizing. For all Raschers, rigorous measurement has been essential to the advancement of civilization since Babylonian times and to the advancement of physical science since the Renaissance and will surely keep with the rhythm of computer and Internet era.

Acknowledgment We would like to thank the anonymous reviewers and our foreign counterparts for suggestions that greatly helped improve the paper.

References

Engelhard, G. Jr. (2013). *Invariant measurement Using Rasch models in the social behavioral, and health sciences*. New York and London: Routledge Taylor & Francis Group.

National Sun Yat-sen University, Taiwan Education Research Association (TERA) and PROMS (2013). Pacific *rim objective measurement symposium*. PROMS Proceeding. unpublished

Trevor, B., & Christine, M. Fox. (2001). *The Preface of Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Paperback ISBN-13: 978-0805842524 and ISBN-10: 0805842527.

Zhang, Q., & Yang, H. (2012). *Pacific rim objective measurement symposium*. German: PROMS Proceeding. the Springer. ISBN 978-3-642-37592-7(ebook) and ISBN 978-3-642-37591-0 (print book).

Yang H., & Miao M. (2014). *A Rasch-Based approach for reading comprehension between CET and GEPT*. @Springer-Verlag Berlin Heidelberg 2015.

Zhang, Q., & Yang, H. (2014). *Pacific rim objective measurement symposium*. German: PROMS Proceeding. the Springer. ISBN 978-3-662-47490-7 (ebook) and 978-3-662-47489-1 (print book).

Zhang Q., & Yang, H. (Eds.). (2014). *Pacific rim objective measurement symposium (PROMS) 2014 conference proceedings*.

Zhang Q., Guoxiong H., & Huifeng M. (2014). *A Rasch-Based approach for listening comprehension between CET and GEPT*. @Springer-Verlag Berlin Heidelberg 2015.

Chapter 3
The Psychometric Properties of the MISSCARE Nursing Tool

I. Blackman, E. Willis, L. Toffoli, J. Henderson, P. Hamilton, C. Verrall, E. Arbery and C. Harvey

3.1 Introduction

Since 2006, US nurse Beatrice Kalisch has explored the relationships among the work environment, patient care demands and staffing issues on nursing outcomes (Kalisch 2006). Subsequently, the MISSCARE (Kalisch and Williams 2009) tool was developed to quantify what types and how frequently nursing care was missed and why omissions occurred. The MISSCARE survey has become one measure in the transactions of nursing, which refers to any aspect of care that is entirely or partially omitted or deferred. The tool comprises two portions: the elements of missed nursing care, containing 24 items where nurse participants are asked to rate how often each care aspect was missed with the options ranging from "rarely," "occasionally," "frequently," and "always" missed. The second component explores the reasons for missed nursing care, with 17 varied reasons for why nursing care was missed within their work area. The scale used offered four options indicating degrees of intensity for why care was missed: if it was a "significant reason," "moderate reason," "minor reason," or "not a reason" for missed care.

I. Blackman (✉) · C. Verrall
School of Nursing & Midwifery, Flinders University, Adelaide, SA, Australia
e-mail: ian.blackman@flinders.edu.au

E. Willis · J. Henderson · E. Arbery
School of Health Sciences, Flinders University, Adelaide, SA, Australia

L. Toffoli
School of Nursing, University of South Australia, Adelaide, SA, Australia

P. Hamilton
Woman's University, Denton, TX, USA

C. Harvey
Eastern Institute of Technology, Hawke's Bay, Napier, New Zealand

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_3

With these measures as a focus, the major aim of this paper is to report findings of a study that determines to what degree nurses' self-rated estimates of the types, frequencies and rationales of missed care (the MISSCARE tool) are psychometrically robust measures using Rasch analysis.

3.2 Background

Rasch analysis assumes that item responses are governed by the person's (self-rated) ability on the underlying trait being estimated. The foundation of Rasch analysis is the modelling of responses based on the person's self-rated ability in relation to the degree of difficulty the items pose, rather than summing total responses, as this causes problems with the final outcomes (Bond and Fox 2007). The Rasch model has been used in a variety of evaluative processes, for example validating or evaluating the psychometric principles of nursing survey instruments (Blackman et al. 2007; Blackman and Hall 2009), health-related concerns (Kelly 2005; Chiu et al. 2006; Hahn et al. 2010; Blackman and Chiveralls 2011; Blackman 2012), assessing the construct validity of test items (Blackman 2009) and determining cut-off points and standards in educational and occupational measurements (Baghaei 2007, 2009).

One reason Rasch analysis is favoured over traditional measurement techniques (called classical test theory) is that much of nursing-related research involves the use of rating scales. Ordinal data as used in Likert-type scales are frequently summed and averaged, and the resultant scores are taken to represent a person's score on a particular outcome such as achievement; however, this practice can lead to unreliable outcomes (Merbitz et al. 1989; Jamieson 2004; Grimby 2012). Such problems occur because it is erroneously assumed that the distance between response choices (e.g. from "strongly disagree" to "agree" or any other category within the scale) represent equal distances (of ability) in the dimension or attitude being estimated.

In reality, the scale categories represent an inherent order of a continuum of the trait being estimated, with numbers given to each category (e.g. 1 or 2 or 3 or 4). These numbers do not indicate the magnitude of difference between the categories and therefore should not be summed or averaged (Cohen et al. 2000). Rasch analysis overcomes this problem with ordinal scales measurement by first transforming Likert data into true interval scales as logarithmic values (known as logits). The resultant interval data can then be used, as it meets the necessary criteria for statistical analysis.

Another feature of Rasch analysis is that participants' ability can be conjointly measured and directly matched to survey item difficulty. This determines to what extent survey item difficulty aligns with the persons' ability, with any mismatch indicating that either the test items were too easy or too difficult for the abilities of the participants (Yates 2005).

The MISSCARE tool has been presented in two past studies, one in the United States of America and the other in Turkey (Kalisch et al. 2011; Duygulu et al.

2012). In both these studies, data were collected for both parts of the survey and results were explored using factor analysis to further confirm the major themes the data were generating. In the US study, the authors maintained that reliability of the two subscales were acceptable and valid, with Cronbach alpha values between 0.69 and 0.85, but it was not clear which parts of the survey these indices referred to. In the later Turkish study, the frequency of missed care (Part A) and rationales for missed care (Part B) had reliability estimates (Cronbach alpha) of 0.95 and 0.67, respectively. These results especially the latter reliability estimates are low, suggesting that the survey items on an individual basis and collectively were not all working to measure the underlying construct that the scale was seeking to examine. A limitation of the reliability of the two MISSCARE studies is that no reliability estimates are made about the participants' responses to the either parts of the surveys. A measure to determine the consistency or the pattern of responses to the survey items is indicated, to explore whether the participants' responses were too erratic or showed little variation in the scale options offered in the survey. Cronbach alpha as a reliability measure does not have the capacity to measure this (Sitjsma 2009). Rasch analysis overcomes this limitation, in that it can confirm if all the survey items are working harmoniously to measure the same underlying (latent) construct being estimated by the surveys (referred to as having unidimensionality), and it can compare participant responses to the modelled patterns (of responses) as predicted by the Rasch model to identify not only reliability of survey item but also reliability of the participant responses (Bond and Fox 2007).

3.3 Methods

3.3.1 Sample/Participants

The survey was undertaken online. Participants were contacted using e-mail to 1600 South Australian Nurses and Midwives Federation (ANMF) members that contained a link to the survey. This was followed up by an advertisement in an electronic newsletter distributed to all ANMF members. The survey was available online for nurses for 2 months from November 1, 2012, to December 31, 2012, and was completed by 289 nurses.

3.4 Instrument

The two Likert-type scales arising from the MISSCARE tool were used to collect data. Type and frequency of missed nursing care over different shifts of work time were also sought with options ranging from "rarely," "occasionally," "frequently,"

Table 3.1 Descriptions of the type of reported missed nursing care according to item number and shift time, as administered to participant surveys

Survey item number	Description of type of nursing care reported as being missed	Type of nursing work shift	Description of type of nursing care reported as being missed	Survey item number
1	Ambulation three times a day as ordered	Early shift	Hand washing	49
2		Late shift		50
3		Night shift		51
4		Weekend shift		52
5	Turning the patient every 2 h	Early shift	Patient discharge planning and education	53
6		Late shift		54
7		Night shift		55
8		Weekend shift		56
9	Feeding patients while food is still warm	Early shift	Bedside glucose monitoring as ordered	57
10		Late shift		58
11		Night shift		59
12		Weekend shift		60
13	Setting up meals for patients who can feed themselves	Early shift	Patient assessments performed each shift	61
14		Late shift		62
15		Night shift		63
16		Weekend shift		64
17	Medications administered within 30 min before or after scheduled time	Early shift	Focussed assessments according to patient condition	65
18		Late shift		66
19		Night shift		67
20		Weekend shift		68
21	Vital signs assessed as ordered	Early shift	IV/Central line care sire and assessment according to hospital policy	69
22		Late shift		70
23		Night shift		71
24		Weekend shift		72
25	Monitoring intake and output	Early shift	Response to call bell/light initiated within 5 min	73
26		Late shift		74
27		Night shift		75
28		Weekend shift		76
29	Full documentation of all necessary data	Early shift		77
30		Late shift	PRN medication requests acted on within 15 min	78
31		Night shift		79
32		Weekend shift		80
33	Patient education about illness, tests and diagnostic tests	Early shift	Assess the effectiveness of medications	81
34		Late shift		82
35		Night shift		83
36		Weekend shift		84
37	Emotional support to patient and/or family	Early shift	Attend interdisciplinary care conferences whenever held	85
38		Late shift		86
39		Night shift		87
40		Weekend shift		88

(continued)

Table 3.1 (continued)

Survey item number	Description of type of nursing care reported as being missed	Type of nursing work shift	Description of type of nursing care reported as being missed	Survey item number
41	Patient bathing/skin care	Early shift	Assist with toileting needs within 5 min of request	89
42		Late shift		90
43		Night shift		91
44		Weekend shift		92
45	Mouth care	Early shift	Skin/wound care	93
46		Late shift		94
47		Night shift		95
48		Weekend shift		96

and "always" missed. A neutral or a not applicable category was intentionally omitted from the scale's design, to maximise data reliability and minimise any episodes of category (threshold) reversal, during the analysis process (Linacre 2002). Table 3.1 highlights the survey items that were used to determine the types and frequencies of missed nursing care.

Additionally, participant responses as to why nursing care was reported to be missed (Kalisch and Williams 2009) were collected together with ratings as to how important participants believed these reasons were to missed nursing care. These estimates were derived by using a second Likert-type scale containing three thresholds as described earlier with options ranging from a "significant reason," "moderate reason," "minor reason," or "not a reason" for missed care. Again a neutral category in this second scale was omitted in favour of maximising a continuum in the intensity of participants' endorsements for why nursing care is missed. The 17 survey items that describe the reasons for reported missed nursing care are listed in Table 3.2, by item number.

3.5 Results

Study data were entered into PASW (version 15.01) file for analysis. Rasch analyses were undertaken using Conquest software developed by the Australian Council of Educational Research (Wu et al. 1998) and Winsteps 3.72 (Linacre 2012).

Table 3.3 indicates that the majority of the respondent nurses were female, aged 45 years and older. The participants were experienced nurses having five or more years nursing experience and employed predominantly in public metropolitan hospitals. The majority of respondent nurses worked 30 or more hours a week.

Table 3.2 Descriptions of the types of reasons why reported nursing care was missed

Item no.	Reason for reported missed nursing care	Item no.	Reason for reported missed nursing care
1	Inadequate number of staff	10	Supplies/equipment not functioning properly when needed
2	Urgent patient situations (e.g. worsening patient condition)	11	Lack of back up support from team members
3	Unexpected rise in patient volume and/or acuity on the ward/Unit	12	Tension or communication breakdowns with other ancillary/support departments
4	Inadequate number of assistive and/or clerical personnel (e.g. care assistants, ward clerks, porters)	13	Tension or communication breakdowns within the nursing team
5	Unbalanced patient assignment	14	Tension or communication breakdowns with the medical staff
6	Medications not available when needed	15	Nursing assistant/carer did not communicate that care was not provided
7	Inadequate handover from previous shift or patient transfers into ward/Unit	16	Nurse/carer assigned to patient off ward/Unit or unavailable
8	Other departments did not provide the care needed (e.g. physiotherapy did not ambulate)	17	Heavy admission and discharge activity
9	Supplies/equipment not available when needed		

3.6 Validity and Reliability of the MISSCARE Tool

3.6.1 Psychometric Qualities of the Survey Scales: Item Fit Statistics of the Two Surveys

As noted earlier, Rasch analysis is based on modelling participant responses. It is essential therefore to explore whether indeed participant responses (to both parts of the missed nursing survey) do conform to the Rasch model's predictions. To do this, the differences between the observed scores (from the surveys) are matched to those as predicted by the Rasch model. Fit indices indicate how well this match fits. Figure 3.1 illustrates whether the survey items used to determine nurses' estimates of the frequency of missed care during a day shift of care are indeed doing exactly that or not. The vertical dotted lines in Fig. 3.1 represent the range of the infit means square values for the survey items and whether the survey items are indeed uni-dimensional (all aligned to measure the same underlying variable), each survey items should fall between these (infit means square) values, less than 0.77 and greater than 1.30. Any items failing to meet these parameters are deemed not to be measuring the same underlying construct as the other items, and they are removed

Table 3.3 Demographic characteristics of nurse respondents

Day and afternoon shift nurses only	N	%
Gender		
Female	261	90
Male	28	10
Age		
Under 25	6	2
25–34	34	12
35–44	57	20
45–54	108	37
55–64	80	27
65 and above	6	2
Years of experience as a nurse		
Less than 2 years	38	13
2–5 years	42	14
5–10 years	45	16
More than 10 years	166	57
Location		
Metropolitan	197	68
Rural	93	32
Setting		
Public	218	75
Private	54	19
Agency	18	6
Number of hours worked		
Less than 30 h/week	95	33
30 h or more/week	195	67
Length of shift		
5–8 h	199	69
9–12 h	76	30
Greater than 12 h	2	1

from any further analysis. Two items fall within this category and are seen to be violating the criterion of unidimensionality of the scale, estimating the frequency of reported missed nursing care. In essence, these two items are measuring some other underlying construct, compared to the other survey items. It is worth noting also that the survey item that has infit values of less than 0.70, (item 25: monitoring intake and output) confirms that responses to this item, in particular, lack the variability of participant responses that the Rasch model was predicting. Conversely, items with fit values above 1.30 (item 85: attends multidisciplinary conferences) suggest that participant responses were too haphazard or more erratic than the mathematical Rasch model was anticipating (Bond and Fox 2007).

In relation to fit statistics associated for the reasons of missed nursing care scale, Fig. 3.2 portrays one item as not fitting the scale. Item 4 (Inadequate number of

```
--------------------------------------------------------------------------------
INFIT
 MNSQ        .53       .63       .77      1.00      1.30      1.60      1.90
--------------+--------+--------+--------+--------+--------+--------+-----------
  item 1                        .         |    *     .
  item 5                        .         |     *   .
  item 9                        .         |    *    .
  item 13                       .         |   *     .
  item 17                       .  *      |         .
  item 21                       .  *      |         .
  item 25                       *         |         .
  item 29                       .       * |         .
  item 33                       . *       |         .
  item 37                       .  *      |         .
  item 41                       .   *     |         .
  item 45                       .   *     |         .
  item 49                       .         |     *   .
  item 53                       .       * |         .
  item 57                       .     *   |         .
  item 61                       .         | *       .
  item 65                       .     *   |         .
  item 69                       .    *    |         .
  item 73                       .         | *       .
  item 77                       .         *         .
  item 81                       .     *   |         .
  item 85                       .         |         .            *
  item 89                       .        *|         .
  item 93                       .         | *       .
================================================================================
```

Fig. 3.1 Fit statistics for survey scale: frequency of missed nursing care (day shift)

```
--------------------------------------------------------------------------------------
INFIT
 MNSQ   .56       .63       .71       .83      1.00      1.20      1.40      1.60
---------+---------+---------+---------+---------+---------+---------+---------+------
  item 1                        .              |       *      .
  item 2                        .              |       *      .
  item 3                        .          *   |              .
  item 4                        .              |              .    *
  item 5                        .              |*             .
  item 6                        .            * |              .
  item 7                        .     *        |              .
  item 8                        .              | *            .
  item 9                        .       *      |              .
  item 10                       .         *    |              .
  item 11                       .        *     |              .
  item 12                       .    *         |              .
  item 13                       .        *     |              .
  item 14                       .            * |              .
  item 15                       .        *     |              .
  item 16                       .            * |              .
  item 17                       .              *              .
======================================================================================
```

Fig. 3.2 Fit statistics for survey scale: reasons for missed nursing care

assistive and/or clerical personnel) is not measuring the same underlying construct (consensus related to why nursing care is missed) as the rest of the survey item set.

As noted earlier, the Cronbach alpha index as a measure of reliability has limitations. The Rasch equivalent of the Cronbach alpha values which explores the reliability of the participants' responses and survey tool items are the Person Separation Index (PSI) and the Item Separation Index (ISI), respectively. These separation indices assess the spread of survey scores both the items and the persons' response patterns across the continuum of the trait being measured which in turn determines the reliability of both parameters. The ISI and PSI values should be in excess of 2 to reflect good reliability of both the scales used in this study (Muis et al. 2009). For the type and frequency of missed nursing care survey, the ISI was 3.91 (reliability 0.94) and the PSI was 3.77 (reliability 0.94), indicating a very good reliability index for the survey items used and stability of the persons' responses to that scale. The attribution scale's reliability indicating why nursing care was missed was also very acceptable with an ISI and PSI of 2.54 (reliability 0.87) and 6.71 (reliability 0.98), respectively. Both scales therefore demonstrate a very acceptable capacity for data replicability should the data be retested (Curtis 2005; de Ayala 2009). Estimates reflecting the reliability of the revised two scales used to measure the frequency of and the reasons for reported missed nursing care confirm their dependability for measurement.

3.6.2 Psychometric Qualities of the Survey Scales: Detecting for Item Bias in the Surveys

Consistency of responses is expected in Rasch analysis, where the survey instruments are expected to function the same way for all respondents undertaking the survey irrespective of the participants' different attributes, e.g such as gender, Hagquist et al. (2009). As described earlier, Rasch analysis is concerned with determining whether invariance in the data obtained from the surveys violates unidimensionality of the construct being measured. Differentiated item functioning (also known as item bias) is one mechanism that explores whether different members or subgroups within the cohort being measured differ markedly from the responses generated by the group as a whole (Bond and Fox 2007). If survey items do differ significantly, it indicates that these items are biased and favour the responses of one subgroup over another group of participants. In such instances, biased survey items need to be removed from the further analysis so as not to produce unreliable outcomes.

In this study, the gender of the nurse participants and whether they worked in a metropolitan or a rural setting were examined as possible sources of differentiated item functioning in both MISSCARE scales. The gender of the nurse was a statistically significant source of invariance (item bias) in the scale which estimated type and frequency of missed nursing care (more than 2 standard deviations from the mean score), however, were not responsible for any item bias in the scale that measured why care was missed. Figure 3.3 shows differentiated item functioning (one item in total) based on the gender groups of nurse participants. It shows that Item 38 (providing emotional support to patient and/or family on a late shift of care) has a greater probability (is more easily) endorsed by the female participants in this group.

Figure. 3.4 and 3.5 shows six item responses are producing unwanted invariance, as these items occur at greater than two standard deviations from the mean. Based on nurses' place of employment and with reference to Fig. 3.4, items 36, 37

```
--------------------------------------------------------------------------------
                  Plot of Standardised Differences
Easier to endorse for     Easier to endorse for male
 Female nursing staff      nursing staff
    -3        -2        -1         0         1         2
 -------+------------+------------+------------+------------+------------+------------
 item 38          .*                 |                 .
=====================================================================
```

Fig. 3.3 Gender item bias (DIF) and frequency of missed nursing care

```
--------------------------------------------------------------------------------
             Plot of Standardised Differences
  Easier to endorse for   Easier to endorse for
metropolitan based        rural based nurses
nurses
    -3      -2      -1       0       1       2       3      4
 -------+---------+---------+--------+---------+---------+---------+--------+---------
 item 36         .              |             *
 item 37         .              |            . *
 item 54         .              |             .         *
=====================================================================
```

Fig. 3.4 Work location, item bias (DIF) and frequency of missed nursing care

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
            Plot of Standardised Differences
Easier to endorse for     Easier to endorse for
for metropolitan           for rural nurses
nurses
    -3        -2        -1         0         1         2
 -------+------------+------------+------------+------------+------------+------------
 item 4            .                 |                 .        *
 item 9          * .                 |                 .
 item 10       *    .                |                 .
=====================================================================
```

Fig. 3.5 Work location, item bias (DIF) and reasons for missed care

and 54 differentiate in favour of the rurally employed nurse participants. In other words, the frequency of missed care for patient education, providing emotional support to patient and/or family and for patient discharge planning, respectively, have a higher probability of being endorsed by rural-based nurses than their metropolitan-based colleagues.

In terms of why nursing care is missed, Fig. 3.5 shows one item (item 4) which is likely to be endorsed by rural nurses with two other items more likely to be endorsed by their metropolitan colleagues (items 9 and 10). Rural nurses are more likely to endorse that care is missed because of inadequate human resources (number of assistive and/or clerical personnel), while metropolitan respondents are more likely to endorse the fact that physical resources are limited (supplies/equipment not available when needed and are not functioning properly when needed), compared to their rural colleagues.

As these seven items are producing unwanted sources of invariance in their respective scales, they have been removed from on-going analysis. As all survey items and participant responses have been screened for validity, reliability and invariance, and it is now possible to estimate the frequency and type of missed nursing care on the same scale (conjointly) to the consensus estimates of participant nurses. Similarly, consensus estimates of the participants can also be conjointly constructed against the different reasons why nursing care is missed.

3.6.3 Psychometric Qualities of the Type and Frequency of Missed Nursing Care Scale: Type of Nursing (Missed) Care and Nurses' Consensus (Frequency) Estimates

The Rasch model is able to place all the survey items for both the frequency of missed care and its rationale as identified by the nurses on hierarchical linear scales. This scaling co-locates each item (indicating each different aspect of nursing care which is missed) to the nurses' capacity to endorse how frequently this nursing care is missed. Figure 3.6 depicts this hierarchy with nurses' ability to endorse the frequency of missed nursing care extending to the left of the dotted vertical line, with type of missed nursing care (called item difficulty) to the right of it. The far right of the figure shows a scale ranging from −5 to +3 which as described previously is the logit (logarithm) scale. The higher the positive value of the scale, the more likely it is that the nursing tasks located at this point are most likely not to be missed: It is more difficult for the nurses to endorse that these nursing activities are not undertaken. Conversely, nursing items located between 0 and −5 show a progressively increasing tendency for these nursing activities to be missed, as it is easier for the participant nurses to endorse that these nursing activities are missed.

At the top of the reported missed nursing care tasks in Fig. 3.6 are Items 49, 50, 51 and 52, and with reference to Table 3.1, this shows that hand washing is the nursing task or skill that is least likely to be missed, across all the four shifts of work surveyed, compared to all the other nursing tasks and skills surveyed. All items occurring below these four nursing skills or tasks on Fig. 3.6 are rated by participants as becoming progressively and more frequently missed, with omitted nursing practice items descending on the hierarchy, culminating in Items 26, 28 and 3 which are located at the bottom of the hierarchy of missed nursing tasks. These three nursing tasks or skills are rated as being most frequently or always missed by respondent nurses. With reference to Table 3.1, items 26 and 28 refer to charting patient's fluid intake and output on late and weekend shifts, respectively, while item 3 is the provision for ambulating patients on night duty. With reference to Fig. 3.6, each surveyed nursing care or task can be reliably identified as being missed or not and conjointly estimated with how frequently that care is missed according to the consensus of participant nurses. Figure 3.6 also identifies each participant's

```
Nurses endorsing
least frequent       | Types of nursing care reported as least missed
missed care
    3                +
                     |
                     |
                     |
                  X  |
                  X  |
    2                +
                    T|
                  X  |
                     |
               XXXX  |
                XXX  |
    1          XXXX  +  49  50  51  52
              XXXXX  |
               XXXX  |  23  43  59  60  66  67  69  70  71
             XXXXXX S|  56  57  58  61  62  63  64  72  79
      XXXXXXXXXXXXX  |  10  11  12  15  31  37  53  77  78  80  81  84  94  95
        XXXXXXXXXXX  |  5  7  9  19  45  47  75  82
    0          XXXX  +M 6  8  17  18  32  34  41  46  48  55  73  74  91
            XXXXXXX  |  13  20  29  30  33  44  76  89  90  92
          XXXXXXXXX  |  4  14  16  39  42  65  68
                XXX  |  1  21  22  24  35  93
          XXXXXXXXX M|  40  96
         XXXXXXXXXX  |  2  27
   -1        XXXXXX  +  83
           XXXXXXXX  |T 26  28
         XXXXXXXXXX  |  3
           XXXXXXXX  |
             XXXXXX  |
             XXXXXX S|
   -2           XXX  +
                XXX  |
               XXXX  |
              XXXXX  |
                  X  |
                  X  |
   -3                +
                    T|
                XXX  |
                     |
                     |
                     |
   -4             X  +
                     |
                     |
                     |
                     |
                     |
   -5         XXXXX  +
Nurses endorsing
most frequent        | Types of nursing care reported as always missed
missed nursing care
```

Fig. 3.6 Frequency and types of reported missed nursing care

likelihood of indicating how frequently each aspect of nursing care is reported as being missed. On the left of the vertical dotted line in that figure are a number of crosses. Each one indicates each participant's interpretations in the frequency of reported missed nursing care. The participant nurse located at the top of the scale is reporting nursing tasks as being least likely to be missed out overall, compared to the rest of the participants. All other respondents (occurring below this nurse's location on the hierarchical scale) are rating nursing care as progressively being missed, with the five participant nurses at the bottom of the scale, indicating that all the surveyed nursing care in their estimation are most likely to be or is always missed.

3.6.4 Psychometric Qualities of the Reasons for Missed Nursing Care Scale: Missed Care Attribution and Nurses' Consensus (Importance) Estimates

Figure 3.7 demonstrates participant nurses' responses to the causes of reported missed nursing care. To the right of the vertical dotted line in hierarchical order are the ranked reasons given by each nurse for why nursing care was missed. Item 16, which is located at the top of all the reasons for missed nursing care in Fig. 3.7, indicates that the availability of an assigned nurse or carer was the least important reason for missed nursing care. It was difficult for nurses to endorse this reason as a major contributor for missed care. Conversely, items 3, 1, 2 and 17, which occur at the lower aspect of the hierarchy of reasons for missed nursing care map, account for the strongest reasons why nursing care is missed.

It was much easier for respondent nurses to endorse these reasons for why nursing care is missed. Unexpected rises in patient volumes (item 3), inadequate staff numbers (item 1), urgent patient situations (item 2) and heavy admission and discharge activity (item 17) attract the strongest consensus as rationales for missed nursing care. On the right of the vertical dotted line in Fig. 3.7 is the frequency of agreement and disagreement about the individual reasons for missed care (each x represents three respondent nurses). The single x at the top of the figure represents nurses who believe that the given causes for missed nursing care as used in the survey are least likely or are not to be responsible for missed nursing care. Conversely, those participant nurses located at the bottom of the map strongly endorse that the surveyed causes for missed nursing care, especially the reasons given adjacent to their position on the scale (that is item 3 in Fig. 3.7) are a major causes for missed nursing care. The majority of nurses agreed that lack up of back up from teams (item 11) were only minor to moderate reasons accounting for why nursing care is missed.

3.7 Discussion

The nurses' self-rated MISSCARE scales are distinctive, as they have been generated to understand the type, frequency and reasons for why nursing care is missed by clinical nurses. By employing Rasch analysis, the two scales were found to have

```
-----------------------------------------------------
Nurses attrib'n: most difficult   Missed care = not a reason
to endorse
  5.0                            |
                                 |
                             X   |
                                 |
                                 |
                                 |
  4.0                            |
                                 |
                                 |
                             X   |
                                 |
                                 |
  3.0                            |
                             X   |
                                 |
                             X   |
                             X   |
                             X   |
  2.0                      XXX   |
                           XXX   |
                            XX   |
                             X   |
                            XX   |
                            XX   |
  1.0                       XX   |16.Assigned nurse or carer not available
                            XX   |12.Tensions in ancillary dep't 13.Tension nurs team
                             X   |15.Missed care was not communicated
                             X   |8.Other dept's did not provide allocated care
                           XXX   |7.Inadequate pt handover 14.Tension with med team
                           XXX   |
                       XXXXXXX   |11.Lack of backup from teams
   .0            XXXXXXXXXXXXX   |
                           XXX   |
                            XX   |5.Unbalanced pt assign't 6.Medic'ns not available
                            XX   |
                             X   |
                            XX   |2.Urgent pt. sit'ns 17.Heavy admit/discharge no's
 -1.0                       XX   |1.Inadequate staff number
                             X   |
                                 |
                             X   |
                             X   |3.Unexpected rise in patient volumes
                             X   |
 -2.0                            |
                                 |
                                 |
                                 |
                                 |
                                 |
 -3.0                            |
Nurse attribution: easiest         Missed care = most significant reason
to endorse
 ------------------------------------------------------------------------------------
                 Each X represents 3 participant nurses.
```

Fig. 3.7 Reasons for missed care: an attribution intensity scale

three items that did not measure the same underlying construct as the rest of the scaled items, which has not been reported previously. With these three items removed, the remaining items on the revised MISSCARE scale fit well with underlying traits being able to be estimated. All remaining items were unidimensional, meaning that nurses' responses to the survey items on both parts of the survey were measuring a single trait defined by the scales offered and not estimating any other variable or underlying factor. This outcome suggests that the internal construct validity and reliability of both the modified MISSCARE scale are quite acceptable.

The four-category response scale as used on the frequency for missed nursing care and similarly for Part B of the MISSCARE tool also demonstrates reliability, as there was no evidence of threshold reversal for any of the categories used in both Likert-type scales. Consequently, the four categories used in the two scales are able to distinguish between the different self-rated capacities for agreement or consensus of the nurses surveyed.

The study's results also demonstrate that the MISSCARE tool performed almost uniformly for nurses of both genders and within different workplaces, except for seven survey items that were found to produce invariance. With those survey items removed, the revised MISSCARE scales do not produce unwanted invariance and are therefore unbiased according to nurses' gender and place of work.

Rasch analysis has also shown that the MISSCARE scales are robust and reliable measures of nurses' consensus for missed nursing care and can provide clinical nurses and managers with a ready mechanism to determine what areas of nursing care, resource allocation and staffing requires minimal supervision and greater support to provide optimal patient care.

The scores given by participant nurses on both the MISSCARE scales provide useful information about how to better understand the relationships between the elements of missed nursing care and the barriers that exists when trying to maximise effective clinical practices.

One major limitation to this study is that the MISSCARE survey may produce invariance across other groups of nurses working in different clinical fields. Additional studies using other nurse cohorts is suggested to ascertain whether the MISSCARE scales perform uniformly across other clinical nurse groups, such as years of expertise.

The self-report tool for estimating episodes for missed nursing care and why it happens can be conjointly measured using Rasch analysis, where consensus estimates of the participants can be directly scaled to different aspects of nursing care and reasons for care omission. Rasch analysis has ensured that all items on both parts of the revised MISSCARE survey are reliable and unidimensional, that the participants' responses fit the parameters of the Rasch model and, finally, that there is no item bias (differentiated item functioning) occurring.

References

Baghaei, P. (2007). Applying Rasch rating scales model to set multiple cut-offs. *Rasch Measurement Transactions, 4*, 1075–1076.

Baghaei, P. (2009). A Rasch-informed standard setting procedure. *Rasch Measurement Transactions, 23*(2), 1214.

Blackman, I. (2009). *Identifying medical student latent variables that influence achievement in graduate-entry medicine*. Germany: VDM Publishing House Ltd.

Blackman, I. (2012). Factors influencing Australian primary production employees' self-efficacy for safety in workplace chemical use. *Workplace Health, 60*(11), 1–8.

Blackman, I., & Hall, M. (2009). Estimating the complexity of applied English language skills to the perceived ability of non-English speaking background student nurses, using Rasch analysis. In B. Matthews & T. Gibbons (Eds.), *The Process of Research in Education A Festschrift in Honour of John P Keeves AM* (pp. 167–183). South Australia, Australia: Shannon Press.

Blackman, I., & Chiveralls, K. (2011). Factors determining the readiness of work-place supervisors to engage in workplace rehabilitation. *Journal of Occupational Rehabilitation, 21* (4), 537–547. doi:10.1007/s10926-011-9297-1.

Blackman, I., Hall, M., & Darmawan, I. (2007). Undergraduate nurse variables that predict academic achievement and clinical competence in nursing. *International Education Journal, 8* (2), 222–236.

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

Chiu, Y., Fritz, S., Light, K., & Velozo, C. (2006). Use of item response analysis to investigate measurement properties and clinical validity of data for the dynamic gait index. *Physical Therapy, 86*(6), 778–787.

Cohen, L., Marion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge Falmer.

Curtis, D. (2005). Comparing classical and contemporary analyses and Rasch measurement. In S. Alagumalai, D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 179–195). The Netherlands: Springer Press.

de Ayala, R. (2009). *The theory and practice of item response theory*. New York: Guildford Press.

Duygulu, S., Kalisch, B., & Terzioglu, F. (2012). The MISSCARE Survey-Turkish: Psychometric properties and findings. *Nursing Economic$*, *30*(1), 29–37.

Grimby, G. (2012). The use of raw scores for ordinal scales; time to end malpractice? *Journal of Rehabilitation Medicine, 44*, 97–98. doi:10.2340/16501977-0938.

Hagquist, C., Bruce, M., & Gustaavsson, J. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies, 46*(3), 380–393. doi:10.1016/j.jnurstu.2008.10.007.

Hahn, E., Cella, D., Bode, R., & Hanrahan, R. (2010). Measuring the social wellbeing in people with chronic illness. *Social Indicators Research, 96*(3), 883–884. doi:10.1007/s11205-009-9484-z.

Jamieson, S. (2004). Likert scale: How to (ab)use them. *Medical Education, 38*(12), 1212–1218. doi:10.1111/j.1365-2929.2004.02012.x.

Kalisch, B. (2006). Missed nursing care: A qualitative study. *Journal of Nursing Care Quality, 21* (4), 306–313.

Kalisch, B., & Williams, R. (2009). Development and psychometric testing of a tool to measure missed nursing care. *The Journal of Nursing Administration, 39*(5), 211–219. doi:10.1097/NNA.0b01e381a23cf5.

Kalisch, B., Tschannen, D., Lee, H., & Friese, C. (2011). Hospital Variation in Missed Nursing care. *American Journal of Medical Quality. 26*(4), 291–299.

Kelly, C. (2005). Commitment to health scale. *Journal of Nursing Measurement, 13*(3), 219–229.

Linacre, J. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2012). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.

Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*, 308–312.

Muis, K., Winne, P., & Edward,s O. (2009). Modern psychometrics for assessing achievement goal orientataion: A Rasch analysis. *British Journal of Educational Psychology*, *79*, 547–576. doi:10.1348/000709908X3834.72.

Sitjsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107. doi:10.1007/s11336-008-9101-0.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *Acer conquest: Generalised item response modelling software*. Victoria, Australia: ACER Press.

Yates, S. (2005). Rasch and attitude scales: Explanatory style. In S. Alagumalai, D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 207–225). The Netherlands: Springer Press.

Chapter 4
Booklet Design for a Longitudinal Study: Measuring Progress in New-Immigrant Children's Mathematics Achievement

Pei-Jung Hsieh

Abstract Because of the increasing number of new-immigrant children in Taiwan, we conducted a large-scale assessment of mathematics achievement at the Grade 4 level in 2012 and performed a follow-up of a sample of panel members in 2014. The objective of the study was to develop a valid instrument in order to measure and investigate the average new-immigrant student's progress between Grade 4 and Grade 6 in mathematical literacy. We compiled 78 selected-response items into a set of 13 booklets. All the test items were selected from the Taiwan Assessment of Student Achievement (TASA 2012), which is the largest nationally representative and continuing assessment in Taiwan. Each item is designed to measure one of the four mathematics content areas, which are number and measurement, geometry, statistics and probability, and algebra. We propose three rules for identifying and selecting appropriate items. First, the number of anchor items is 39 (50 %), providing the basis for equating scores on different grades. Second, we designated as high priorities Grade 4 items with a difficulty parameter larger than −0.5 and Grade 6 items with a parameter between −1 and 0.5. Third, the proportions of each content area should be similar to those listed in the General Guidelines of Grades 1–9 Curriculum for Mathematics. We applied a three-parameter logistic model and used PARSCALE to estimate the item parameters. After these procedures, each test block contained six items, and each booklet comprised four blocks based on the balanced incomplete block design. The mean difficulty of the items in each block ranged from 0.053 to 0.153. Consequently, the mean difficulty of the items in each booklet ranged from 0.083 to 0.128. The distributions of items across the content areas were 64.10 %, 19.23 %, 5.13 %, and 11.54 %, mostly corresponding to the curriculum framework. This study demonstrated that a detailed consideration of the percentage of anchor items, the range of item difficulties, and the distribution of content areas can be useful for constructing a measurement tool in a longitudinal study.

Keywords Balanced incomplete block design · Achievement growth · Panel study

P.-J. Hsieh (✉)
National Academy for Educational Research, New Taipei City, Taiwan
e-mail: pjh@mail.naer.edu.tw

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_4

4.1 Introduction

With the continuous decrease in the fertility rate of Taiwanese citizens, the proportion of new-immigrant children among primary school students continues to increase every year. In the 2004–2005 school-year, there were 1,883,533 pupils, and this number dropped to 1,297,120 in the 2013–2014 school year, a decrease of 31.13 %. However, the number of new-immigrant pupils increased from 40,907 in the 2004–2005 academic year to 157,431 in the 2013–2014 academic year, an increase from 2.17 to 12.14 %. The highest number of these children had mothers with Chinese nationality, followed by mothers with Vietnamese and Indonesian nationalities (Ministry of Education 2014). Following school admission, additional attention should be paid to education-related topics such as academic performance, parent–teacher communication, and adjustment to school by children born in transnational marriage families.

The National Academy for Educational Research conducted a large-scale investigation for fourth and sixth-grade students in 2012. The results showed that children from new-immigrant families in the fourth and sixth grades performed significantly worse in mathematics than children from nonimmigrant families. However, the difference in performance seemed to decline with age. In addition, growth in academic performance could vary among children based on the differences in their family socioeconomic status (Wang et al. 2012). Besides the correlation between family socioeconomic status and performance in mathematics, which could change with time, the relationship between psychological variables and academic achievement is a subject of concern in the educational sector. To further understand the effects of the environmental context of new immigrant children and time on their performance in mathematics and to find more convincing evidence, we resampled the 2012 fourth-grade subjects to conduct a long-term longitudinal study on the same population of new-immigrant children in 2014.

To understand the growth trend in new-immigrant children's performance in mathematics, to investigate the casual pathways of the children's performance in mathematics, and to analyze the effect of education policies on performance in mathematics, this study developed reliable and valid evaluation tools to accurately measure the growth in the mathematics performance of new immigrant children in fourth to sixth grades.

4.2 Content Standards

We developed a mathematics framework based on the General Guidelines of Grades 1–9 Curriculum for Elementary and Junior High School Education. A mathematics curriculum equips students with an understanding of the basic concepts of figures, shapes, and quantities as well as the ability to calculate and organize and to apply such knowledge and skills in daily life. It also enables comprehending the principles of reasoning and problem solving, the ability to elaborate clearly on mathematics-related concepts, and making appropriate connections among materials and contents between this and other learning areas. The mathematics curriculum for Grades 1–9 is divided into four stages: Stage 1 begins in Grade 1 and ends in Grade 3; Stage 2 begins in Grade 4 and ends in Grade 5; Stage 3 begins in Grade 6 and ends in Grade 7; and Stage 4 begins in Grade 8 and ends in Grade 9 (Ministry of Education 2006).

4.3 Test Item Bank

The mathematics items used in the 2012 large-scale assessment of new-immigrant children were constructed by the TASA, which is the largest nationally representative and continuing assessor of Taiwanese students' knowledge and skill-sets in five subject areas. Assessments are conducted periodically in mathematics, Mandarin, English, science, and social science. TASA assessments began in 2005. (For a detailed description of the TASA assessment plan, see Table 4.1)

TASA classifies mathematics assessment questions into two dimensions: content area and mathematical complexity. Each question is designed to measure one of four mathematics content areas: (a) number and measurement, (b) geometry, (c) statistics and probability, and (d) algebra. Moreover, items are classified according to three types of mathematical abilities: conceptual understanding, procedural knowledge, and problem solving.

The distribution of items among the various mathematical content areas and mathematical complexities reflects the relative proportion of the mathematics curriculum. In 2012, there were 65 selected-response test items and 13 constructed-response test items for Grade 4 as well as Grade 6 students. Table 4.2 lists the distribution of selected-response items, and Table 4.3 shows the

Table 4.1 TASA assessment schedule

Year/grade	Mathematics	Mandarin	English	Science	Social science
2014	11	11	11	11	11
2013	8	8	8	8	8
2012	4, 6	4, 6	6	4, 6	6
2011	11	11	11	11	11
2010	8	8	8	8	8
2009	4, 6	4, 6	6	4, 6	6
2007	4, 6, 8, 11	4, 6, 8, 11	4, 6, 8, 11	4, 6, 8, 11	6, 8, 11
2006	4, 6, 8, 11	4, 6, 8, 11	4, 6, 8, 11	4, 6, 8, 11	6, 8, 11
2005	6	6	6		

Table 4.2 Distribution of selected-response items for Grade 4 in 2012

Content areas	Types of mathematical abilities			Number of items
	Conceptual understanding	Procedural knowledge	Problem solving	
Number and measurement	17	12	14	43
Geometry	4	4	4	12
Statistics and probability	1	1	2	4
Algebra	2	2	2	6
	24	19	22	65

Table 4.3 Distribution of constructed-response items for Grade 4 in 2012

Content areas	Types of mathematical abilities			Number of items
	Conceptual understanding	Procedural knowledge	Problem solving	
Number and measurement	0	1	7	8
Geometry	0	1	2	3
Statistics and probability	0	0	1	1
Algebra	0	0	1	1
	0	2	11	13

Table 4.4 Distribution of selected-response items for Grade 6 in 2012

Content areas	Types of mathematical abilities			Number of items
	Conceptual understanding	Procedural knowledge	Problem solving	
Number and measurement	14	10	16	40
Geometry	5	3	5	13
Statistics and probability	0	0	3	3
Algebra	2	2	5	9
	21	15	29	65

distribution of constructed-response test items for Grade 4 in 2012. In addition, Table 4.4 details the distribution of selected-response items, and Table 4.5 displays the distribution of constructed-response test items for Grade 6 in 2012. In order to link the test scale scores between Grade 4 and Grade 6, we embedded 17 common selected-response items into both item pools.

TASA and our 2012 large-scale assessment of new-immigrant children use a balanced incomplete block (BIB) design to assign blocks or groups of selected-response cognitive items to student booklets. A BIB design satisfies four

Table 4.5 Distribution of constructed-response items for Grade 6 in 2012

Content areas	Types of mathematical abilities			Number of items
	Conceptual understanding	Procedural knowledge	Problem solving	
Number and measurement	0	0	6	6
Geometry	0	1	1	2
Statistics and probability	0	1	1	2
Algebra	0	1	2	3
	0	3	10	13

Table 4.6 TASA balanced incomplete block booklet design

Booklet version	Position			
	Position 1 cognitive block	Position 2 cognitive block	Position 3 cognitive block	Position 4 cognitive block
S1	M3	M2	M1	M10
S2	M4	M5	M10	M6
S3	M7	M10	M8	M9
S4	M10	M11	M12	M13
S5	M5	M1	M9	M11
S6	M2	M6	M11	M7
S7	M11	M3	M4	M8
S8	M1	M4	M7	M12
S9	M8	M12	M5	M2
S10	M12	M9	M6	M3
S11	M13	M7	M3	M5
S12	M9	M13	M2	M4
S13	M6	M8	M13	M1

conditions: every treatment in the booklet design is covered at most only once in a booklet; every treatment in the booklet design appears with equal frequency across all booklets; every booklet has an identical length, containing the same number of clusters; and every pair of treatments in the booklet design occurs together in the booklets with equal frequency (Frey et al. 2009). We assigned 65 selected-response items into a set of 13 booklets, and each booklet comprised four blocks (i.e., every student response to 20 items). The BIB booklet design (Table 4.6) enabled TASA and our 2012 large-scale assessment of new-immigrant children to sample a sufficient number of students to obtain precise results for each test.

Table 4.7 Distribution of selected-response items for new-immigrant children in 2014

Content areas	Types of mathematical abilities			Number of items
	Conceptual understanding	Procedural knowledge	Problem solving	
Number and measurement	10 + 10	8 + 3	11 + 8	29 + 21
Geometry	1 + 4	2 + 2	2 + 4	5 + 10
Statistics and probability	1 + 0	0 + 0	1 + 2	2 + 2
Algebra	0 + 2	1 + 1	2 + 3	3 + 6
	12 + 16	11 + 6	16 + 17	39 + 39

Note The number before the + sign represents the number of Grade 4 items, whereas the number after the + sign represents the number of Grade 6 items

4.4 Method

Because of budget limitations, we chose only selected-response items from 2012 for the 2014 booklets. We propose three rules for identifying and selecting appropriate items. First, the percentage of anchor items must be 50 %, providing the basis for equating scores on different grades. Second, the Item Response Theory model of TASA is constrained to have a mean ability of zero, and thus, the TASA sample represents the whole population of Taiwanese students with a higher mathematics ability compared with new-immigrant children. Therefore, the Grade 4 items with a difficulty parameter larger than −0.5 and the Grade 6 items with a parameter between −1 and 0.5 are considered high priority. Third, the proportion of each content area should be similar to that in the General Guidelines of Grades 1–9 Curriculum for Mathematics. PARSCALE 4 is used to estimate the item parameters.

4.5 Results

Totally, we selected 78 selected-response items. The distribution of items is listed in Table 4.7. The item distribution across the content areas was 64.10, 19.23, 5.13, and 11.54 %, which is relatively similar to the original distribution of the Grade 4 and Grade 6 items. Following the 2012 BIB design, we compiled all of the items into a set of 13 booklets, and each booklet comprised four blocks (i.e., every new-immigrant student response to 24 selected-response items in 40 min). The mean difficulty in the items on each block ranged from 0.053 to 0.153, and consequently, in each booklet it ranged from 0.083 to 0.128 (Table 4.8).

Table 4.8 The mean difficulty of the items on each booklet

Booklet version	Mean difficulty on each block				Mean difficulty on each booklet
	Position 1 cognitive block	Position 2 cognitive block	Position 3 cognitive block	Position 4 cognitive block	
S1	M3 0.153	M2 0.100	M1 0.152	M10 0.093	0.125
S2	M4 0.092	M5 0.128	M10 0.093	M6 0.053	0.092
S3	M7 0.092	M10 0.093	M8 0.097	M9 0.143	0.106
S4	M10 0.093	M11 0.087	M12 0.132	M13 0.117	0.107
S5	M5 0.128	M1 0.152	M9 0.143	M11 0.087	0.128
S6	M2 0.100	M6 0.053	M11 0.087	M7 0.092	0.083
S7	M11 0.087	M3 0.153	M4 0.092	M8 0.097	0.107
S8	M1 0.152	M4 0.092	M7 0.092	M12 0.132	0.117
S9	M8 0.097	M12 0.132	M5 0.128	M2 0.100	0.114
S10	M12 0.132	M9 0.143	M6 0.053	M3 0.153	0.120
S11	M13 0.117	M7 0.092	M3 0.153	M5 0.128	0.123
S12	M9 0.143	M13 0.117	M2 0.100	M4 0.092	0.113
S13	M6 0.053	M8 0.097	M13 0.117	M1 0.152	0.105

4.6 Conclusion

This study demonstrated that a detailed consideration for the percentage of anchor items, the range of item difficulties, and the distribution of content areas can be useful for constructing a measurement tool in a longitudinal study.

Acknowledgments This research was supported by the Ministry of Science and Technology (MOST 102-2511-S-656-005) and the National Academy for Educational Research.

References

Frey, A., Hartig, J., & Rupp, A. A. (2009). An ncme instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53. doi:10.1111/j.1745-3992.2009.00154.x.

Ministry of Education. (2006). *General guidelines of grades 1-9 curriculum for elementary and junior high school education*. Taipei, Taiwan: Ministry of Education.

Ministry of Education. (2014). *Overview of new immigrant's students (2013-2014 school year)*. Taipei, Taiwan: Ministry of Education.

Wang, R.-J., Kuo, K.-B., & Cheng, C.-M. (2012). *Factors influencing the academic performance of new-immigrant children in Taiwan: Individuals, families, and schools*. Taipei, Taiwan: National Academy for Educational Research.

Chapter 5
Indicators of Integration of Islamic Values in Human Knowledge

Nik A. Hisham, Kamal J. Badrasawi and Noor Lide Abu Kassim

Abstract The International Islamic University Malaysia (IIUM) is committed to the Integration of Islamic values in the Human Knowledge (IOHK), which is the core of its vision and mission. One of the main concerns of IOHK is to determine indicators to evaluate the implementation of IHOK at IIUM and other Malaysian public institutions of higher learning. This study aims to explore the psychometric properties of a set of indicators measuring the integration of Islamic values in the curriculum at IIUM using the Rasch Measurement Model. This study utilized the survey method, with a 65-item questionnaire developed based on the literature review and data taken from focus group discussions. The items were divided into seven subconstructs: B*elief in IOHK* (BI), *Content of IOHK* (CO), *Teaching and Learning Process* (TL), *Evaluation* (EV), *Purpose of IOHK* (PS), *Product* (PR), and *Student Improvement* (SI). A total of 324 academic staff randomly selected from various faculties at IIUM completed the questionnaire. The statistical software Winsteps, version 3.72.1, was used to conduct the analysis of the polytomous data collected. The Rasch analysis showed that overall item and person reliability values were very high (0.99 and 0.96, respectively), with item separation 10.51 and person separation 4.81. All items had positive point measure correlation coefficients, with few items below 0.3. Fit statistics estimates showed that all items were within the recommended acceptable range (0.5–1.50), except 5 items on BI subconstruct. Variance explained by measures, indicated useful measurement (52.4 %) with the variance explained by the first contrast in the residual less than 10 %. The item and person means measures were well matched, 0.0 and 0.05 logit, respectively. There were no gaps in the middle of the item distribution, but quite wide gaps were seen at the upper and lower ends of the scale. Some items were overlapping; however, they

N.A. Hisham (✉) · K.J. Badrasawi
Faculty of Education-International Islamic University Malaysia, Kuala Lumpur, Malaysia
e-mail: nikahmad@iium.edu.my

N.L.A. Kassim
Faculty of Dentistry—International Islamic University Malaysia, Kuala Lumpur, Malaysia

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_5

measure different aspects. The instrument as it stands provides useful measures. Nonetheless, the items related to BI and SI and the gaps at the opposite ends of the scale require qualitative investigation.

Keywords Rasch analysis · Knowledge integration · Higher institutions

5.1 Introduction

Curriculum is an important component in education, as it determines how the education process should take place. In general, it comprises the interlinked elements of the learning process: aims, content, methods, and evaluation (Taba 1962). The curriculum should be designed toward reality; i.e., it should consider the cultural, social, ideological, spiritual, philosophical, and psychological dimensions of society. It should also consider the theories of learning styles and human development. In the Islamic community, Islam is the religion and the way of life (Al-Faruqi 1982). Thus, Islam should be the guiding framework in the curriculum design and the main source of reference.

Since its inception in 1983, the International Islamic University Malaysia (IIUM) has been committed to the process of integration of Islamic principles and values with the modern fields of knowledge. Although this process of integration into the university curriculum has been actively promoted (Sskemanya et al. 2007), little effort has been made to evaluate its success. To this end, a set of indicators to measure the integration of Islamic principles and values in the curriculum have been developed; however, its psychometric properties have not been examined. The current study, therefore, aims to examine the psychometric properties of these indicators through the use of the Rasch Measurement Model (RMM).

5.2 Research Method

The study employed a quantitative method of data collection and analysis. A survey was used to gather information on the curriculum integration done by IIUM academic staff. The self-developed survey questionnaire was based on the literature review and data taken from focus group discussions. It consisted of 65 items and was divided into two sections. Section A elicited demographic information about the respondents, including gender, faculty, nationality, post, and year of service. Section B consisted of seven subconstructs, namely, *Belief in Integration of Knowledge* (B; 7 items), *Content* (CO; 8 items), *Teaching & Learning Process* (TL;

19 items), *Evaluation* (EV; 9 items), *Purpose* (PS; 6 items), *Product* (PR; 12 items), and S*tudent Improvement* (SI; 5 items). All of the items for sections B to SI were measured using the 5-point Likert-type scale.

A sample size of 324 academic staff from 12 faculties at IIUM was selected for this study. The RMM was used to analyze the data. The statistical software Winsteps, version 3.72.1, was used to conduct the Rasch analyses of the polytomous data (Linacre 2011). RMM is widely used in most research areas, as it has the ability to extend the evidence of construct validity, explore construct unidimensionality, and produce estimates of item and person score reliability (Bond and Fox 2007; Wright and Stone 1979). In other words, RMM ensures the validity of items by (a) examining item polarity, fit statistics, and unidimensionality and (b) checking the consistency of the items with the purpose of the study through investigating reliability indices for both items and persons.

5.3 Results and Discussion

Academic staff perception on the integration of Islamic principles and values was analyzed using RMM, and the analysis outputs are depicted in Figures and Tables. First, the adequacy of the overall scale of integration of Islamic values was examined in three aspects: consistency with purpose of measurement (reliability and separation), validity of items (item polarity and item fit), unidimensionality of the measured construct and targeting and items ordering. Second, the examination of the subconstructs of the integration scale was also conducted.

5.4 Overall Scale of Integration of Islamic Values in the Human Knowledge

5.4.1 Consistency with Purpose of Measurement

Table 5.1 shows that a total of 324 persons with 65 items were measured. The values of reliability for item difficulty and person ability are very high (0.99 logit and 0.96 logit respectively). This suggests that the ordering of item difficulty and person ability is highly replicable with other similar samples (Bond and Fox 2007; Wright and Stone 1979). The items separation index is 10.51, implying that the items can be divided into 11 levels, while the separation index for persons is 4.81, indicating that persons can be divided into 5 levels. The separation index value greater than 2 is considered as productive (Bond and Fox 2007; Linacre 2011).

Table 5.1 Difficulty measure, fit statistics, item correlation, reliability and separation for all items

No.	Item label	Item measures	(SE)	INFIT MNSQ	OUTFIT MNSQ	PT-measure CORR
1	B1	−2.29	0.10	2.15	3.14	0.09
2	B2	−0.88	0.07	1.88	2.78	0.16
3	B3	−0.77	0.06	1.34	1.56	0.27
4	B4	−1.7	0.08	1.79	2.38	0.08
5	B5	0.27	0.06	2.04	2.53	0.06
6	B6	0.01	0.06	1.72	2.11	0.12
7	B7	−1.68	0.08	1.40	2.23	0.12
8	CO1	−0.48	0.06	0.61	0.68	0.59
9	CO2	0.04	0.06	1.28	1.57	0.27
10	CO3	−0.87	0.07	1.10	1.32	0.19
11	CO4	0.54	0.06	1.07	1.12	0.52
12	CO5	0.32	0.06	1.45	1.78	0.18
13	CO6	0.53	0.06	1.56	1.77	0.28
14	CO7	0.36	0.06	0.86	0.96	0.54
15	CO8	0.23	0.06	0.79	0.89	0.59
16	TL1	−0.67	0.06	0.69	0.67	0.59
17	TL2	−0.6	0.06	0.62	0.61	0.6
18	TL3	0.17	0.06	0.71	0.70	0.7
19	TL4	−0.73	0.06	0.61	0.59	0.61
20	TL5	−0.6	0.06	0.59	0.57	0.64
21	TL6	0	0.06	0.63	0.62	0.69
22	TL7	0.01	0.06	0.63	0.63	0.68
23	TL8	−0.12	0.06	0.70	0.67	0.68
24	TL9	−0.26	0.06	0.60	0.58	0.67
25	TL10	−0.07	0.06	0.76	0.73	0.66
26	TL11	0.38	0.06	0.74	0.73	0.7
27	TL12	0.5	0.06	0.82	0.81	0.68
28	TL13	0.78	0.06	0.99	1.04	0.57
29	TL14	0.49	0.06	0.94	0.93	0.64
30	TL15	0.55	0.06	0.76	0.74	0.7
31	TL16	0.42	0.06	0.86	0.85	0.65
32	TL17	0.8	0.06	0.84	0.80	0.68
33	TL18	0.37	0.06	0.74	0.73	0.68
34	TL19	0.15	0.06	0.68	0.67	0.69
35	EV1	0.02	0.06	0.62	0.61	0.69
36	EV2	0.24	0.06	0.71	0.72	0.69
37	EV3	0.52	0.06	1.05	1.01	0.63
38	EV4	0.78	0.06	0.83	0.78	0.69
39	EV5	0.11	0.06	0.84	0.83	0.64

(continued)

Table 5.1 (continued)

No.	Item label	Item measures	(SE)	INFIT MNSQ	OUTFIT MNSQ	PT-measure CORR
40	EV6	−0.23	0.06	0.67	0.69	0.66
41	EV7	0.54	0.06	0.89	0.85	0.68
42	EV8	0.18	0.06	0.77	0.76	0.68
43	EV9	0.5	0.06	0.91	0.91	0.64
44	PS1	0.37	0.06	1.11	1.09	0.6
45	PS2	0.55	0.06	1.00	1.01	0.58
46	PS3	0.34	0.06	1.24	1.26	0.51
47	PS4	0.66	0.06	1.14	1.14	0.58
48	PS5	−0.21	0.06	1.25	1.32	0.43
49	PR1	0.34	0.06	1.17	1.12	0.62
50	PR2	0.55	0.06	0.99	0.96	0.66
51	PR3	0.46	0.06	1.08	1.05	0.66
52	PR4	0.55	0.06	1.14	1.09	0.66
53	PR5	0.83	0.06	1.09	1.03	0.7
54	PR6	0.79	0.06	1.08	1.03	0.65
55	PR7	0.66	0.06	0.84	0.80	0.7
56	PR8	0.58	0.06	0.94	0.89	0.68
57	PR9	0.58	0.06	1.06	1.01	0.67
58	PR10	0.56	0.06	1.18	1.12	0.64
59	PR11	0.1	0.06	0.91	0.88	0.65
60	PR12	0.01	0.06	0.84	0.82	0.65
61	SI1	−1.08	0.07	1.07	1.01	0.45
62	SI2	−1.11	0.07	1.04	0.96	0.46
63	SI3	−1.19	0.07	1.15	1.13	0.44
64	SI4	−1.22	0.07	1.12	1.03	0.44
65	SI5	−0.96	0.07	1.09	1.11	0.45
Means		0.0	0.06	1.01	1.08	
Item reliability/ Item separation		0.99 10.51	Person reliability/ Person separation		0.96 4.81	

5.4.2 *Validity of Items*

5.4.2.1 Item Polarity

Item polarity, represented by the point–measure correlation coefficient (PTMEA CORR), provides information on the extent to which all items are working in the same direction to measure the construct being examined (Bond and Fox 2007). Relatively high and positive values (0.3–0.8) are wanted (Bond and Fox 2007;

Linacre 2011). Table 5.1 shows that all items have positive point correlation coefficients, but few items are below 0.3 showing that these items were not effectively discriminating persons with different levels of ability. There is a high possibility that correlation values get higher if misfit persons were deleted. Nevertheless, all items are measuring the construct in the same direction.

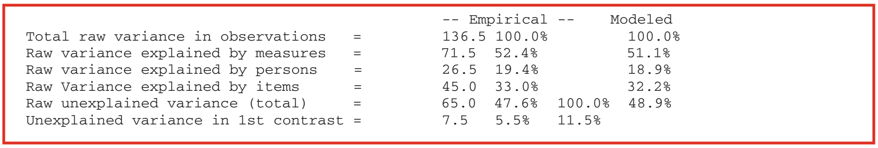
5.4.2.2 Fit Statistics

Item fit statistics (infit MNSQ and outfit MNSQ statistics) are always examined to ensure that the items are contributing meaningfully to the measurement of the construct (Bond and Fox 2007; Linacre 2011). The recommended acceptable range for infit MNSQ and outfit MNSQ fit statistics for rating scale is MNSQ ≥ 0.50 to MNSQ ≤ 1.50 (Bond and Fox 2007; Linacre 2011). Items within this range are considered productive (Bond and Fox 2007). Table 5.1 reveals that all items show good overall fit of the data to Rasch Model. Only six items (B1, B2, B4, B5, B6 and C6) show poor fit (INFIT and OUTFIT > 1.5 logit), and only four items (B3, B7, CO2 and CO5) have OUTFIT > 1.5 logit.

Having deleted some of most misfit persons, only the items on B, namely (B1, B2, B4, B5), remained misfitting (INFIT and OUTFIT > 1.5 logit). These items measuring B could be measuring a construct different from aspects related to the curriculum (i.e., curriculum content, teaching and learning approach, evaluation, purpose, production and student improvement). This is supported by the Rasch analyses for each individual subconstruct. It is important to maintain that the scale unidimensionality was not violated as shown in Fig. 5.1. All the misfit items, therefore, were retained for the aforementioned reasons. However, it would be informative to analyze the first subconstruct separately and the other subconstructs together.

5.4.2.3 Construct Unidimensionality

In RMM, the items must measure a single unidimensional construct (Bond and Fox 2007). The principal component analysis of residuals was used to test the unidimensionality of the measured construct. Having deleted some of the misfit persons, the variance explained by measures indicated a useful measurement (52.4 %), with the variance explained by the first contrast in the residual less than 10 % (about 5.5 %) as shown in Fig. 5.1 (Linacre 2011).



```
                                              -- Empirical --    Modeled
Total raw variance in observations   =        136.5 100.0%         100.0%
Raw variance explained by measures   =        71.5  52.4%          51.1%
Raw variance explained by persons    =        26.5  19.4%          18.9%
Raw Variance explained by items      =        45.0  33.0%          32.2%
Raw unexplained variance (total)     =        65.0  47.6%  100.0%  48.9%
Unexplained variance in 1st contrast =        7.5   5.5%   11.5%
```

Fig. 5.1 Standardized residual variance (in Eigenvalue units)

5.4.2.4 Person and Item Distributions

Figure 5.2 (Item-Person Map) shows the distribution of all items and persons on one logit scale. The item difficulty measure spanned from −2.29 logits to 0.83 logit, while the person ability measure spanned from −2.50 to 6.95 logits. There are no wide gaps in the item distribution, except at the upper and lower positions of the scale. Nonetheless, the person and item distributions are well matched (item mean = 0.0 logit and person mean = 0.5 logit respectively). Looking at the overlapping items, it is found that most of them are measuring either different subconstructs or different aspects of a subconstruct. Figure 5.2 also shows that SI is the easiest to be agreed upon by the respondents.

5.4.3 Validity of Persons' Responses

Finally, validity of persons' responses was also examined. Fifty-six (17 %) persons showed poor fit (INFIT and OUTFIT > 1.5 logit). Figure. 5.2 and 5.3 show the most misfitting responses strings to items (Figs. 4 and 5).

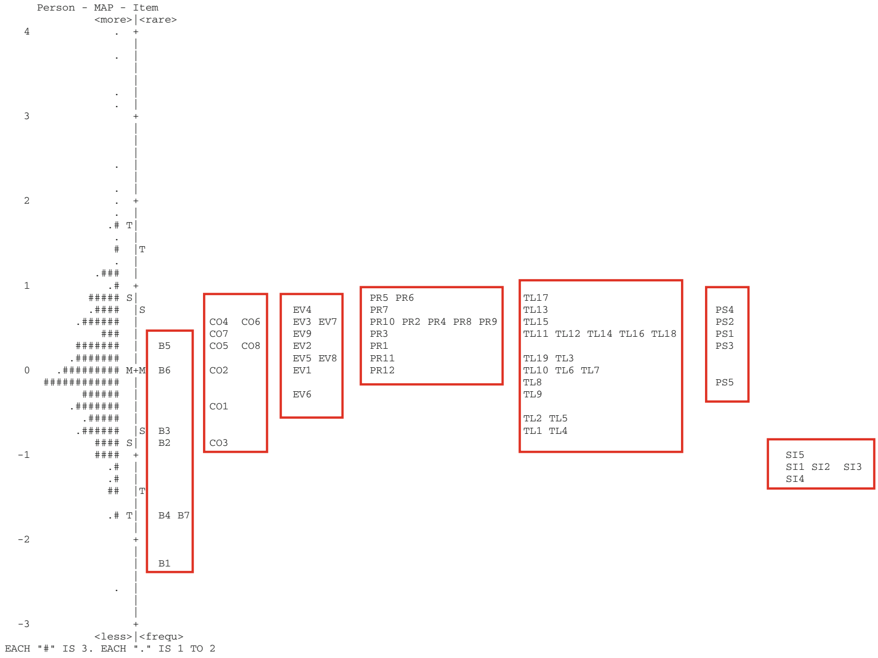
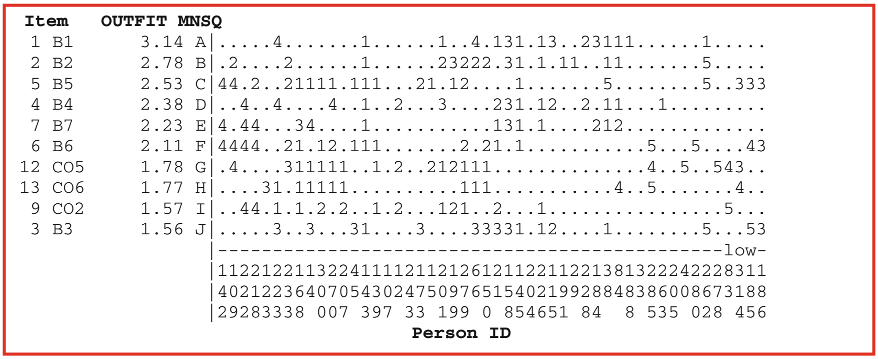
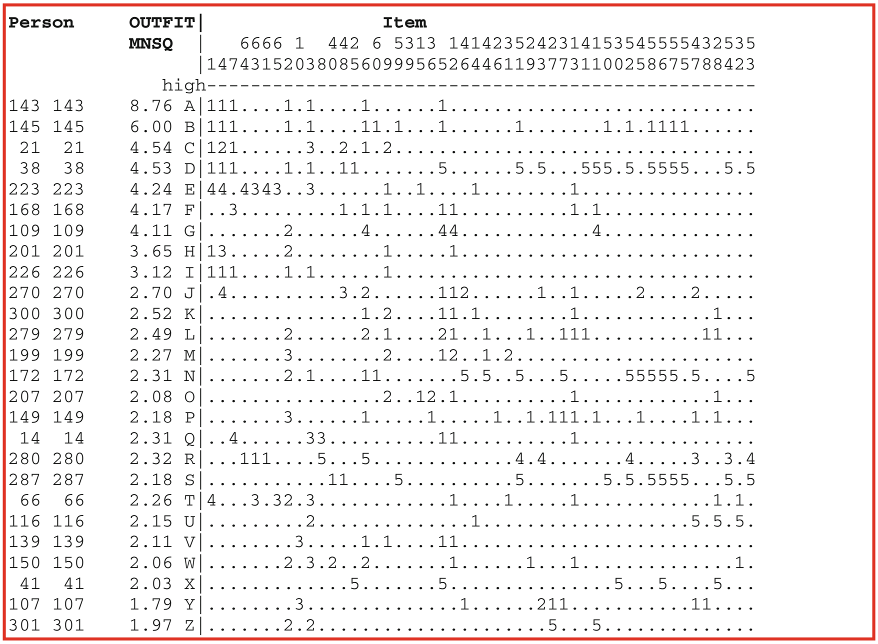


Fig. 5.2 Item-person map: distribution of items for all the seven subconstructs of integration



```
 Item   OUTFIT MNSQ
  1 B1       3.14 A|.....4.......1......1..4.131.13..23111......1.....
  2 B2       2.78 B|.2....2......1......23222.31.1.11..11.......5.....
  5 B5       2.53 C|44.2..21111.111...21.12....1.......5........5..333
  4 B4       2.38 D|..4..4....4..1..2...3....231.12..2.11...1.........
  7 B7       2.23 E|4.44...34....1...........131.1....212.............
  6 B6       2.11 F|4444..21.12.111.......2.21.1...........5...5....43
 12 CO5      1.78 G|.4....311111..1.2..212111..............4..5..543..
 13 CO6      1.77 H|....31.11111..........111...........4..5.......4..
  9 CO2      1.57 I|..44.1.1.2.2..1.2...121..2...1................5...
  3 B3       1.56 J|.....3..3...31....3....33331.12....1........5...53
                   |-----------------------------------------------low-
                   |11221221132241111211212612112211221381322242228311
                   |40212236407054302475097651540219928848386008673188
                   |29283338 007 397 33 199 0 854651 84  8 535 028 456
                                     Person ID
```

Fig. 5.3 Rasch analysis output showing most misfitting response strings



```
Person     OUTFIT|                Item
           MNSQ  |   6666 1  442 6 5313 1414235242314153545555432535
                 |14743152038085609995652644611937731100258675788423
              high--------------------------------------------------
143 143     8.76 A|111....1.1....1......1............................
145 145     6.00 B|111....1.1....11.1...1......1.......1.1.1111......
 21  21     4.54 C|121......3..2.1.2.................................
 38  38     4.53 D|111....1.1..11.......5......5.5...555.5.5555...5.5
223 223     4.24 E|44.4343..3......1..1....1........1................
168 168     4.17 F|..3.........1.1.1....11..........1.1..............
109 109     4.11 G|.......2......4......44............4..............
201 201     3.65 H|13.....2........1.....1...........................
226 226     3.12 I|111....1.1......1.................................
270 270     2.70 J|.4..........3.2......112......1..1.....2....2.....
300 300     2.52 K|..............1.2....11.1........1............1...
279 279     2.49 L|.......2......2.1....21..1...1..111..........11...
199 199     2.27 M|.......3........2....12..1.2......................
172 172     2.31 N|.......2.1....11.......5.5..5...5.....55555.5....5
207 207     2.08 O|................2..12.1..........1............1...
149 149     2.18 P|.......3......1.....1.....1..1.111.1...1....1.1...
 14  14     2.31 Q|..4......33..........11..........1................
280 280     2.32 R|...111....5...5.............4.4.......4.....3..3.4
287 287     2.18 S|...........11....5..........5........5.5.5555...5.5
 66  66     2.26 T|4...3.32.3............1....1.....1............1.1.
116 116     2.15 U|.........2..............1...................5.5.5.
139 139     2.11 V|........3.....1.1....11...........................
150 150     2.06 W|.......2.3.2..2.......1......1...1..............1.
 41  41     2.03 X|.............5.......5...............5...5....5...
107 107     1.79 Y|........3..............1......211...........11....
301 301     1.97 Z|.......2.2.....................5...5..............
```

Fig. 5.4 Rasch analysis output showing most misfitting response strings

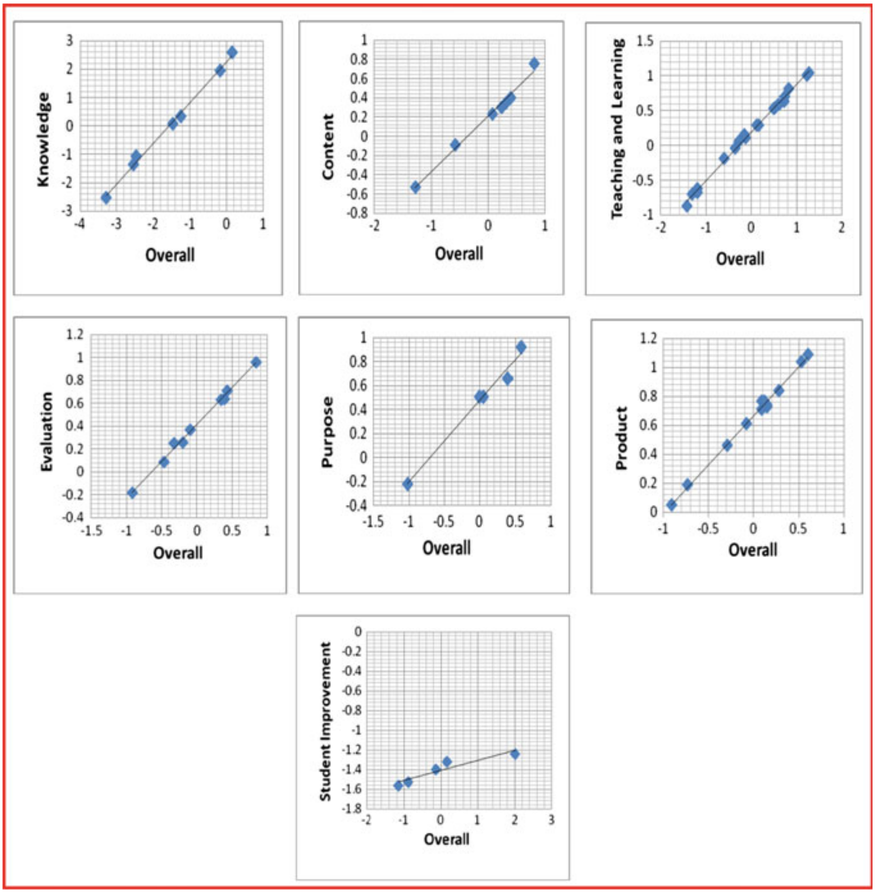


Fig. 5.5 Scatterplots of item difficulty measures analyzed separately and together

5.5 Examination of the Subconstructs of Integration of Islamic Values in the Human Knowledge Scale

5.5.1 Reliability and Separation Index

Table 5.2 shows item and person reliability and separation measures for all the subconstructs of the integration scale. After deletion of misfitting persons, the reliability item measures ranged from 0.95 to 0.99, (B: 0.99, CO: 0.98, TL: 0.99, EV: 0.98, PS: 0.98, PR: 0.96, and SI: 0.95), while the person reliability measures ranged from 0.80 to 0.94, (B: 0.80, CO: 0.83, TL: 0.94, EV: 0.88, PS: 0.83, PR: 0.91, and SI 0.92). Bond and Fox (2007) assert that the reliability value greater than 0.8 is acceptable. Table 5.2 shows that all the person and item separation values are

Table 5.2 Reliability and separation analysis for each sub-construct

Construct	ID item	Item deleted	Item measure		Person measure	
			Reliability	Separation	Reliability	Separation
Belief	B1-B7	–	0.99	13. 53	0.80	2.01
Content	CO1-C08	–	0.99	8.56	0.80	2.01
Teaching and Learning	TL1-TL19	–	0.99	10.32	0.94	3.92
Evaluation	EV1-EV9	–	0.98	6.60	0.88	2.76
Purpose	PS1-PS5	–	0.98	6.67	0.83	2.18
Production	PR1-PS12	–	0.96	5.20	0.90	2.96
Student Improvement	SI1-SI5	–	0.95	4.25	0.92	3.34

greater than 2, the recommended acceptable value given by Bond and Fox (2007). Moreover, deletion of a misfitting item (CO3) led to better measures on the CO subconstruct.

5.5.2 *Item Polarity and Fit Statistics*

Table. 5.3 shows the point measure correlation coefficient (PTMEA CORR) and fit statistics (infit MNSQ and outfit MNSQ) for all items on all subconstructs. All items have positive point correlation coefficient (PTMEA CORR), and infit MNSQ values are within the recommended acceptable range (0.5–1.50), except item CO3. Hence, it could be said that all items of each subconstruct were able to discriminate persons with different levels of ability and working in the same direction to measure the intended construct. However, the point correlation coefficients (PTMEA CORR) for the items on the last subconstruct SI have high values and closer to 1 (0.92–0.97). The respondents might see the items on this subconstruct have the same reference, SI_2 (personality), SI_5 (appearance), SI_3 (morality); and SI_4 (attitude). So it is recommended that these items be further examined qualitatively. For the fit statistics, the infit MNSQ showed that only one item, (CO3, 1.52 logit) is above the recommended acceptable range. This item deals with a controversial issue among staff, i.e., the difficulty of integration of knowledge into the curriculum content. The item was deleted, and its deletion increased the item correlation coefficients and person reliability and separation.

Table 5.3 Item polarity and item fit statistics of all sub-constructs

Item	Belief (B) PTME INFIT OUTFIT CORR			Content (CO) PTME INFIT OTFIT CORR			Teaching (TL) PTME INFIT OUTFIT CORR			Evaluation (EV) PTME INFIT OUTFIT CORR			Person (PS) PTME INFIT OUTFIT CORR			Production (P) PTME INFIT OUTFIT CORR			St. Improvement (SI) PTME INFIT OUTFIT CORR		
Item1	0.55	1.06	0.87	0.68	0.73	0.74	0.66	1.04	1.00	0.78	0.90	0.94	0.79	1.17	1.11	0.82	0.94	0.94	0.96	0.88	0.38
Item2	0.66	1.22	1.17	0.57	0.98	0.98	0.66	1.01	0.97	0.78	0.97	0.97	0.84	0.71	0.68	0.77	1.46	1.54	0.97	0.63	0.24
Item3	0.69	0.86	0.92	0.22	1.52	1.54	0.74	1.10	1.10	0.76	1.19	1.21	0.84	0.74	0.69	0.83	0.79	0.79	0.97	0.75	0.29
Item4	0.61	0.96	0.93	0.73	0.95	0.95	0.68	0.89	0.86	0.79	0.93	0.87	0.81	0.97	0.91	0.83	0.84	0.83	0.97	0.78	0.41
Item5	0.67	1.06	1.05	0.60	0.99	1.00	0.70	0.84	0.83	0.77	1.05	1.05	0.72	1.39	1.43	0.82	0.93	0.85	0.92	1.47	1.42
Item6	0.70	0.85	0.85	0.68	1.11	1.10	0.77	0.81	0.80	0.77	0.94	0.93				0.84	0.76	0.78			
Item7	0.55	0.97	1.04	0.72	0.73	0.72	0.75	0.84	0.91	0.78	1.06	1.06				0.82	0.89	0.97			
Item8							0.73	1.01	1.00	0.81	0.82	0.79				0.84	0.74	0.70			
Item9							0.78	0.81	0.78	0.76	1.05	0.1.08				0.84	0.74	0.80			
Item10							0.74	0.99	0.95							0.81	1.00	1.02			
Item11							0.79	0.86	0.85							0.77	1.36	1.30			
Item12							0.77	1.03	1.04							0.75	1.43	1.49			
Item13							0.69	1.42	1.73												
Item14							0.72	1.36	1.35												
Item15							0.80	0.84	0.83												
Item16							0.74	1.10	1.18												
Item17							0.77	1.09	1.04												
Item18							0.77	0.92	0.89												
Item 19							0.78	0.77	0.77												

5.6 Item Distribution and Difficulty Measures of Each Subconstruct

The distributions of the items on each subconstruct were examined to see the possible ordering of these items. It is found that items on each subconstructs varied in their difficulty measures. Table 5.4 shows the most difficult and easiest items on each subconstruct as endorsed by the respondents.

5.6.1 *Scatterplot*

Scatterplots were produced to examine the order of the difficulty measures of the items on each subconstruct when analyzed together and separately. The plots indicate that the ordering of item difficulty measures of the subconstructs in the individual analyses match the difficulty measures in the overall analysis. However, the last subconstruct, SI, needs further qualitative investigation.

Table 5.4 Most endorsed and least endorsed items on each sub-construct

Sub-construct	Most endorsed	Difficulty measure	Least endorsed	Difficulty measure
Belief (B)	B1	−2.53	B5	2.59
	An important mission of IIUM		*An overemphasized mission*	
Content (CO)	CO1	−1.28	CO4	0.81
	Integrates IOK		*Is all about IOK*	
Teaching & Learning (TL)	TL4	−1.43	TL17	1.27
	Giving examples		*Games*	
Evaluation (EV)	EV6	−0.92	EV4	0.84
	Class presentation		*Colloquium*	
Purpose (PS)	PS5	−1.02	PS4	0.58
	My colleague		*External expert*	
Production (PR)	PR12	−0.91	PR5	0.6
	Student assignment		*Books*	
Student Improvement (SI)	SI 4	−1.15	SI5	2.01
	Attitude		*Appearance*	

5.7 Conclusion

The psychometric properties of the indicators to measure the integration of Islamic principles and values in the curriculum at IIUM have been examined through the use of the Rasch Measurement Model. These indicators could provide a useful measurement. Nonetheless, the items related to B and SI subconstructs and the gaps at the opposite ends of the scale require further qualitative investigation.

References

Al-Faruqi, I. R. (1982). *Islamization of knowledge: General principles of work plan*. Herndon: IIIT.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Lawrence.

Linacre, J. M. (2011). *A user's guide to Winsteps & Ministep Rasch-Model computer programs. Program Manual 3.72.1*. http://www.winsteps.com/a/winsteps.pdf.

Ssekamanya, S. A., Suhailah, H., & Nik Ahmad, H. (2007). The experience of Islamization of Knowledge at the International Islamic University Malaysia: Successes and challenges. New intellectual horizon in education.

Taba, H. (1962). *Curriculum development: Theory and practice*. New York: Harcourt, Brace & World Inc.

Wright, B. D., & Stone, M. H. (1979). *Best design test: A handbook for Rasch measurement*. Chicago: Mesa Press.

Chapter 6
A Rasch Measure of Test Takers' Attitude Toward the Versant English Test

Jinsong Fan and Trevor Bond

Abstract Despite the steadily increasing application of Rasch modeling in human sciences, few attempts have been made to use the Rasch measurement model to analyze Likert-type scale survey data in the field of language testing. This study investigated test takers' attitude toward the Versant English Test (VET), a fully automated spoken English test using the Rating Scale Model (RSM) in Rasch measurement. Based on previous research, attitude in this study was conceptualized as a psychological construct consisting of three components: beliefs, opinions, and emotions. A 21-item questionnaire was designed to collect the data from 125 VET test takers. The collected data were then subjected to Rasch analysis including item statistics, reliability and separation indexes, category structure, Rasch factor analysis, and differential item functioning. The results indicated that the questionnaire was essentially unidimensional, tapping into a single attitudinal construct. The principle of measurement invariance held across the relevant subsamples. The findings of this study have implications for the VET provider in light of the further improvement and promotion of the test. Furthermore, this study also has methodological implications for researchers in the field of language testing.

Keywords Attitude · ESL · The Versant English test · Rasch model · Rating scale analysis

6.1 Background

Attitude refers to "a mental and neural state of readiness organized through experience, exerting a directive or dynamic influence upon individual's response to all objects and situations in which it is related" (Allport 1971, p. 13). In the realm of

J. Fan (✉)
College of Foreign Languages and Literature, Fudan University, Shanghai, China
e-mail: jinsongfan@fudan.edu.cn

T. Bond
College of Arts, Society and Education, James Cook University, Townsville, Australia
e-mail: trevor.bond@jcu.edu.au

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_6

social psychology, the study of attitude has a long history, and researchers have generally come to the consensus that attitude is a multifaceted psychological construct that consists of three interrelated components: cognitive, affective, and conative (e.g., Eagly and Chaiken 1993). Compared to the research of attitude in social psychology, the interest in attitude in language studies is comparatively recent, primarily in the domain of second language acquisition. To date, numerous studies have demonstrated that more positive attitudes of learners toward the target language or the target language culture have a beneficial impact on their language learning and achievement (e.g., Brown 2000; Gardner 1985). Despite the widely recognized role of attitude in language learning, there is no evidence that it has been adequately researched in the field of language testing (Murray et al. 2012).

The prime objective of this study is to use the Rasch measurement model to investigate test takers' attitude toward the Versant English Test (VET), a fully automated spoken English test developed by Pearson. Although some empirical studies of test takers' attitude toward language tests have been reported (e.g., Elder et al. 2002; Fan and Ji 2014; Zhao and Cheng 2010), none of them was focused on a spoken English test, let alone a fully automated one. Methodologically speaking, all previous attitudinal studies in language testing, to the best of our knowledge, have adopted the Classical Test Theory (CTT) for data analysis (e.g., Fan and Ji 2014; Murray et al. 2012; Zhao and Cheng 2010). Given the limitations of the CTT in processing Likert-type scale attitudinal data (e.g., Bond and Fox 2007; Embreston and Reise 2000), this study adopted Rasch measurement, claimed as "the only technique generally available for constructing measures in the human sciences" (Bond and Fox 2007, p. 263), in investigating test takers' attitude toward the VET. It is expected that this study could provide VET stakeholders with credible evidence as to how test takers view this automated spoken English test, thereby paving the way for the future improvement and promotion of this test. Meanwhile, although the application of Rasch modeling has been increasing steadily across the human sciences (e.g., Bond and Fox 2007; Cavanagh and Waugh 2011), in language testing, this technique is currently limited to test validation research (e.g., Bachman 2000; McNamara and Knoch 2012). Few attempts, to the best of our knowledge, have been made to apply the Rasch measurement models to process Likert-type scale survey data in the field of language testing.

6.2 Attitude in Language Testing

In language testing, attitude is often considered as akin to face validity, a concept which is defined as "surface credibility and public acceptability of a test" (Ingram 1977, p. 18). Since face validity is not based a statistical model but on the subjective evaluation of lay people (e.g., students, teachers), it has been often dismissed as unscientific and irrelevant, especially in the earlier research literature (e.g., Bachman 1990; Stevenson 1985). However, this ostensibly plausible view about attitude is not tenable because attitude as a psychological construct entails a broader scope of inquiry which often, if not always, subsumes face validity. For example, affective factors such as test-taking motivation and anxiety are frequently investigated in attitudinal studies but are seldom included in research that pivots on face validity. Furthermore, in response to the view that face validity is unscientific and irrelevant, convincing counterarguments have been raised by language testing researchers. Alderson et al. (1995), for example, argued that if test takers consider a test to be face valid, "they are more likely to perform to the best of their ability on that test and respond appropriately to items" (p. 173). In a similar vein, Karelitz (2014) argued that public opinion of a test should be studied routinely throughout the life cycle of a test because negative public views "create a unique threat to the existence of a test" (p. 4). These arguments for face validity resonate with Messick's (1989) view that test takers' attitude should be considered as a crucial source of construct validity. Shohamy (2001) discussed this issue from the power perspective, arguing that involving test stakeholders such as test takers in test development and validation helps to promote power sharing and fairness in language testing.

A review of the relevant literature reveals that although the tripartite division of attitude (i.e. cognitive, affective, and conative, see Eagly and Chaiken 1993) has been extensively recognized in language studies (e.g., Baker 1988, 1992; Ladegaard 2000), the attitude construct has been conceptualized in manifold ways in previous research, as manifested by the different terms used to represent test takers' attitude, including, for example, "reactions" (Elder et al. 2002), "feedback" (Brown 1993), "views" (Han et al. 2004; Wu 2008), and "psychological factors" (Jin and Cheng 2013). Accompanying the terminological inconsistency are the different operationalizations of this construct in previous studies. In Rasti's (2009) study of Iranian test takers' attitude toward the IELTS, for example, attitude was operationalized as test takers' views on the four components in the test battery, including listening, reading, writing, and speaking. Fan and Ji (2014) adopted a similar definition in their investigation of test takers' attitude to the Fudan English Test (FET), an in-house English proficiency test. In their study, attitude was operationalized as test takers' perceptions of the effectiveness or usefulness of the different aspects of the FET such as test design, administration, and washback. Compared to Rasti (2009) and Fan and Ji (2014), the framework used by Murray and associates (2012) better mirrored the understanding of this construct in the realm of social psychology. In their study, they identified three interrelated factors that represented test takers' attitude toward the Professional English Assessment for Teachers (PEAT) in Australia: *beliefs* (that a proposition is or is not true), *opinions* (that an actual or hypothetical action should or should not happen), and *emotions* (corresponding to the affective component in the tripartite division of attitude). Since this conceptualization best represents the theoretical understanding of attitude in the realm of social psychology, it was adopted in the present study in our investigation of test takers' attitude toward the VET.

6.3 The Versant English Test

The VET, known more widely as the PhonePass test before assuming its current name in 2005, is a fully automated spoken English test that applies state-of-the-art technology in the assessment of the language abilities of nonnative speakers (Pearson 2008; see also Chun 2006). According to the test provider, the VET has been widely used by academic institutions, corporations, and government agencies throughout the world to evaluate the ability of students, staff, or officers to understand spoken English and to express themselves clearly and appropriately in English (Pearson 2008, p. 3). The VET may be administered either over the telephone or through computer and takes approximately 15 min to complete. As a fully automated spoken English test, the VET system can analyze test takers' responses and report their test scores within minutes of the completion of the test. Test administrators and score users can view and print out test scores from a password-protected website.

The test results that test takers receive include a numeric composite score, ranging from 20 to 80, and four diagnostic subscores in Sentence Mastery, Vocabulary, Fluency, and Pronunciation, all ranging from 20 to 80. The composite score is the weighted total of the four diagnostic scores and can be converted to the Common European Framework of Reference (CEFR) global scale (Council of Europe 2001) to facilitate test score interpretations and use, although the details as to how the alignment was conducted are not given either on the VET website or on the test description and validation summary. For example, those scoring 69–78 on the VET are Proficient Users, located at the C1 level of the CEFR scale (Pearson 2008, p. 14).

The construct measured in the VET is the "facility" in spoken language which is defined as "the ability to understand the spoken language on everyday topics and to speak appropriately in response at native-like conversational pace in an intelligible form of the language" (Bernstein et al. 2010, p. 358). More specifically, the VET assesses the "automaticity" with which test takers can respond to the test tasks. "Automaticity" refers to "the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate responses without conscious attention to the linguistic code" (Pearson 2008, pp. 8–9). The construct in the VET is operationalized through the six tasks in the test: Reading, Repeat, Short Answer Questions, Sentence Builds, Story Retelling, and Open Questions. The test format of the VET is presented in Table 6.1 together with a brief description of each task in the VET and the number of items in each part. Of the 63 items in the VET, 57 responses are currently used in the automatic scoring, excluding the two items in Open Questions and each first item in Reading, Repeat, Short Answer Questions, and Sentence Builds (Pearson 2008).

Several strands of validity evidence have been collected to support test score interpretations and use which are primarily in the two areas of construct representation and concurrent validity (e.g., Bernstein et al. 2010; Pearson 2008). In terms of construct representation, Bernstein et al. (2010) explained in detail the concept of

Table 6.1 Test format and content of the VET

Item	Type	Task description	Number of items
(1)	Reading	Test takers read printed, numbered sentences, one at a time, in the requested order	8
(2)	Repeat	Test takers repeat sentences verbatim	16
(3)	Short answer questions	Test takers listen to spoken questions in English and answer each question with a single word or short phase	24
(4)	Sentence builds	Test takers rearrange three short phrases presented in a random order into a sentence	10
(5)	Story retelling	Test takers listen to a story and describe what happened in their own words	3
(6)	Open questions	Test takers present their views or opinions after listening to a question in English	2

"facility" in spoken language and articulated how this construct contributes directly to test scores and is operationalized through the six tasks in the VET, as shown in Table 6.1. In addition to construct representation, concurrent validation efforts have lent strong support to the validity of the VET. For example, in two studies comparing human rating and machine rating, correlation coefficients were reported at 0.81–0.86 (n = 151, Present-Thomas and Van Moere 2009) and 0.77 (n = 130, Farhady 2008), indicating high levels of agreement in test takers' performance on the VET and human tests. Thus, as Bernstein et al. (2010, p. 374) concluded, the facility scoring implements an empirically derived quantitative model of listening and speaking performance at different levels of L2 proficiency, and the psychometric data suggest that facility is an important component of effective speech communication or oral proficiency. Despite the multiple strands of evidence that have been collected in support of the validity of the VET, no studies have ever used the Rasch measurement models to investigate how test takers view this fully automated spoken English test. Test takers' attitude, as explicitly recommended by Messick (1989), provides crucial insights into the validity of a test. This study is therefore intended to fill in this research gap.

6.4 The Present Study

The prime objective of this study is to investigate test takers' attitude toward the VET, using the Rasch measurement model. The limitations of the CTT in processing Likert-type scale data have been extensively discussed in Rasch literature (e.g., Bond and Fox 2007; Cavanagh and Waugh 2011). The principal disadvantage of the CTT lies in that it relies on sample statistics to derive scale estimates. Consequently, different scale properties (e.g., item-total correlations, Cronbach's alphas) may be yielded with different samples, thus making it difficult to generalize the research findings (Embreston and Reise 2000; Oon and Subramaniam 2011).

Second, in analyzing the Likert-type scale survey data, the CTT assumes that the scale is linear, all items have the same impact, and that the distance between any two adjacent categories is equal. However, as Reid (2006, p. 12) pointed out, "there is no way of knowing whether the scale in an individual attitude question is linear, equally spaced." In a similar vein, Bond and Fox (2007, p. 101) argued that the CTT approach disregarded the subjective nature of the data by "making unwarranted assumptions about their meaning." The relative value of each response category across all items is treated as being the same, and the unit increases across the rating scale are given equal value. The CTT approach in analyzing Likert-type scale data, as Bond and Fox (ibid.) continued to argue, is therefore "both counterintuitive and mathematically inappropriate."

To overcome the limitations of the CTT, this study adopted the Rasch measurement model which is based on the assumption that the probability of any person being successful on any test item is governed by item difficulty (D) and person ability (B; Rasch 1960, 1980). The instrument that was used to collect the data in this study was an attitude questionnaire, developed on the basis of the conceptualization of attitude articulated by Murray et al. (2012). Specifically, this study seeks to investigate the following questions:

RQ1 Does the questionnaire measure a single underlying construct, i.e., test takers' attitude toward the VET?

RQ2 According to the Rasch analyses of the questionnaire items, what are test takers' attitudes toward the VET?

RQ3 Do the items in the questionnaire show the property of measurement invariance across the relevant subsamples in this study?

6.5 Method

6.5.1 Participants

The participants in this study are 125 students from a research university in east China, all studying for their bachelor's degree at the time when this study was conducted. For all participants, Chinese is their first language. Their age ranged from 17 to 23 years old (Mean = 20). Among the 125 participants, 76 (60.8 %) were females and 49 (39.2 %) were males. They came from different academic backgrounds, with 58 (46.4 %) studying subjects in the domain of humanities and 67 (53.6 %) in science.

6.5.2 Instruments

An attitude questionnaire was developed on the basis of the theoretical framework articulated by Murray and associates (2012). In this framework, attitude is posited to be a psychological construct consisting of three components: beliefs, opinions, and emotions. The initial draft of the questionnaire was intended to be as comprehensive as possible, containing a total of 35 items, all on a six-point Likert-type scale of agreement (from 1 to 6: strongly disagree—disagree—slightly disagree—slightly agree—agree—strongly agree). The questionnaire was piloted on a group of 54 students within the same university where the main study was to be conducted. Based on the feedback from students and some experienced researchers, as well as some initial Rasch analyses, a 21-item questionnaire was produced for this research. The 21 items were broadly categorized into 4 content areas: (1) test takers' attitude toward the design of the VET, (2) test takers' perceived difficulty of the tasks in the VET (3) test takers' perceived interest of the tasks in the VET, and (4) test-taking motivation. At the end of the questionnaire, participants were requested to provide their biodata, including their gender, age, and academic background.

6.5.3 Data Collection

Due to practical constraints, convenience sampling, rather than strictly stratified sampling, was employed in this study. Two months before the study, we sent e-mails to 300 prospective participants, calling for their participation in this study. Thanks to the generous support of the VET provider, all participants were exempt from the charges of taking the VET. It turned out that 125 students took the test and completed the questionnaires, achieving a response rate of 41.7 %. The rather low response rate was understandable, since currently the VET is not widely known to Chinese university students. The VET was administered to the participants in two university language laboratories in April 2014, with the aid of a proctor from the VET provider. After the administration of the VET, each participant signed the informed consent form and completed the questionnaires.

6.5.4 Data Analysis

In this study, the RSM (Andrich 1978) in Rasch measurement was used to analyze the questionnaire data. The RSM is an extension of the simple (i.e., the dichotomous) Rasch model (Rasch 1960, 1980), and is routinely used to calibrate and examine the quality of response categories in Likert-type scales. The mathematical expression of this model is presented as:

$$\log\left(P_{nij}/P_{ni(j-1)}\right) = B_n - D_i - F_j,$$

where P_{nij} and $P_{ni(j-1)}$ refer to the probability of a person n of ability B_n being observed as responding to category j or lower category $j - 1$, respectively, of a rating scale on a particular item i of difficulty D_i, with F_j the threshold calibration which is held as constant across all items in the rating scale (Bond and Fox 2007).

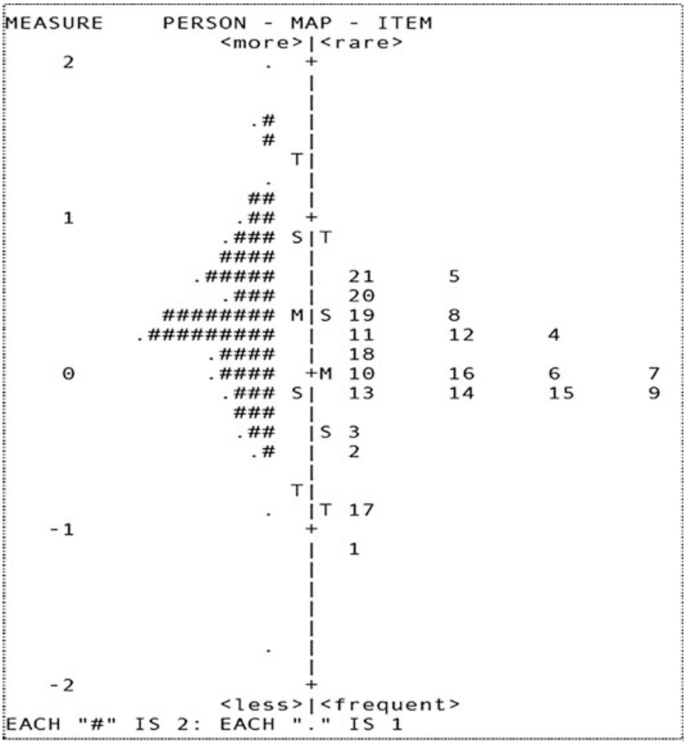
To address the three research questions, we first examined all the item statistics including the item-person map, point-measure (PTMEA) correlations, and infit and outfit mean squares (MnSq) with a view to investigating whether the items in the questionnaire fit the expectations of the Rasch model. At the same time, the Rasch-modeled item measures provided crucial evidence as to how test takers perceived the VET. Second, we examined the reliability estimates and separation indices as well as the response category structure to examine further the quality of the questionnaire. Third, Rasch factor analysis was performed to investigate whether additional dimensions existed in the variance unexplained, or unmodeled, by the primary Rasch measure. Finally, we performed Differential Item Functioning (DIF) to check whether the principle of measurement invariance held across appropriate subsamples in this study: Test takers were divided into two subsamples according to gender and academic background, respectively. All analyses in this study were performed using Rasch software Winsteps 3.81.0 (Linacre 2012).

6.6 Results and Discussion

6.6.1 *The Item-Person Map*

A distinctive advantage of Rasch analysis lies in that it can graphically illustrate the locations of items and persons on the interval-level measurement scale. The item-person map contains a lot of basic information that is central to Rasch measurement (Bond and Fox 2007). Since measures of items and persons are calibrated on the same scale, it can reveal the relationships between item endorsement and person agreement simultaneously. It can also determine whether the item difficulties are appropriate (i.e., well targeted) for the targeted sample. The item-person map of the 21 items with 125 participants is presented in Fig. 6.1.

In Fig. 6.1, "M" represents mean for persons and items, "S" represents one standard deviation (SD) away from the mean, and "T" represents two SD away from the mean. On the left side of the figure, persons (i.e. test takers) were arranged in the order of their endorsement of the questionnaire items. Those located at the upper end agreed most with the items whereas those at the lower end agreed least. The right side of the figure depicts the items arranged in endorsability order. Items at the top were most difficult to be endorsed by the test takers, while items toward the bottom were easier to be endorsed. As indicated by this figure, the items were reasonably well targeted at the persons with item measures ranging from −1.07 to 0.63 and person



```
MEASURE      PERSON - MAP - ITEM
                   <more>|<rare>
   2                  .  +
                         |
                         |
                    .#   |
                     #   |
                       T |
                     .   |
                    ##   |
   1               .##   +
                  .### S |T
                  ####   |
                .#####   |  21     5
                  .###   |  20
              ######## M |S 19     8
             .#########  |  11     12     4
                 .####   |  18
   0             .####   +M 10     16     6      7
                  .### S |  13     14     15     9
                   ###   |
                   .##   |S 3
                    .#   |  2
                         |
                       T |
                     .   |T 17
  -1                     +
                         |  1
                         |
                         |
                         |
                         |
                     .   |
                         |
  -2                     +
                   <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Fig. 6.1 Item-person map of the 21 items with 125 participants

measures from −1.73 to 1.98. In other words, the cluster of persons was located more or less opposite to the cluster of items (Bond and Fox 2007). However, this figure also reveals that no items were targeted at the persons at the upper end of the scale. This indicates that this scale was too easy for this sample, and more difficult questions were therefore necessary to raise the "ceiling" of the scale.

6.6.2 Item Statistics

Item statistics from Rasch analysis are presented in Table 6.2 including item measures, infit and outfit MnSq statistics and standardized Z values (Zstd), and PTMEA correlations. Following the procedures suggested by Linacre (2012) to investigate data dimensionality, we first of all examined PTMEA correlations; negative values would indicate that items were improperly scored or did not function as intended. All items in the questionnaire demonstrated positive correlation coefficients. Furthermore, all items except Item 5 exhibited moderate to strong correlations (0.33–0.68). The correlation coefficient of Item 5 was 0.24

Table 6.2 Item statistics for Rasch analysis (n = 125)

Items	Measure	Error	Infit		Outfit		PTMEA correlation
			MnSq	Zstd	MnSq	Zstd	
Test design							
1. The demo before the test helps me to get used to the test	−1.07	0.11	1.18	1.30	1.08	0.70	0.33
2. I like the design of the VET	−0.53	0.09	0.87	−1.10	0.83	−1.40	0.52
3. I feel the tasks in the VET are engaging	−0.32	0.09	0.87	−1.10	0.84	−1.30	0.51
4. I think the VET can accurately reflect my spoken English ability	0.22	0.08	0.91	−0.07	0.90	−0.80	0.41
Perceived difficulty of the VET							
5. I believe the VET as a whole is a difficult test	0.61	0.08	1.15	1.30	1.16	1.40	0.24
6. I think ‘Repeat’ is difficult to me	0.02	0.09	0.95	−0.40	1.04	0.40	0.38
7. I think ‘Questions’ is difficult to me	0.06	0.09	0.80	−1.70	0.79	−1.90	0.51
8. I think ‘Sentence Builds’ is difficult to me	0.32	0.08	1.03	0.30	1.03	0.40	0.38
9. I think ‘Story Telling’ is difficult to me	−0.10	0.09	1.09	0.70	1.06	0.50	0.50
10. I think ‘Open Questions’ is difficult to me	0.00	0.09	1.20	1.60	1.20	1.60	0.33
Perceived interest of the tasks							
11. I think ‘Reading’ is interesting to me	0.19	0.08	1.13	1.10	1.13	1.10	0.52
12. I think ‘Repeat’ is interesting to me	0.25	0.08	0.88	−1.00	0.89	-0.90	0.57
13. I like the design of ‘Questions’	−0.15	0.09	0.96	−0.30	0.96	-0.30	0.68
14. I enjoyed working on ‘Sentence Builds’	−0.15	0.09	1.24	1.90	1.19	1.50	0.66
15. I enjoyed working on ‘Open Questions’	−0.15	0.09	1.13	1.00	1.10	0.80	0.57
Test-taking motivation							
16. I was looking forward to taking the VET	−0.01	0.09	1.02	0.20	0.99	0.00	0.46
17. I took the VET to assess my English ability	−0.86	0.10	0.90	−0.70	0.85	−1.10	0.50
18. I took the VET because I like English	0.10	0.09	1.00	0.10	1.04	0.40	0.34

(continued)

Table 6.2 (continued)

Items	Measure	Error	Infit		Outfit		PTMEA correlation
			MnSq	Zstd	MnSq	Zstd	
19. I took the VET because its scores are widely recognized	0.43	0.08	0.83	−1.50	0.83	−1.50	0.51
20. I took the VET to obtain its certificate	0.50	0.08	0.99	−0.10	0.99	0.00	0.46
21. I took the VET to seek better employment	0.63	0.08	0.99	0.00	1.00	0.00	0.44

Note The items in this table were translated loosely from the original questionnaire which was presented in Chinese; the content was somewhat adjusted for the sake of brevity

which was somewhat below the criterion of 0.30. These results suggest that the items in the questionnaire were functioning in the same direction.

Following the examination of PTMEA correlations, the second step to investigate data dimensionality, according to Linacre (2012), is to examine item fit statistics that assess the extent to which the data have been modeled by the strict mathematical expectations of the Rasch model. Items that do not fit the Rasch model generally do not define the same common construct. To investigate the fit of the items, infit and outfit statistics were used which adopt slightly different techniques for assessing an item's fit to the Rasch model (Bond and Fox 2007, p. 57). The infit statistic (weighted) gives relatively more weight to the performances of persons closer to the item value, while the outfit statistic is not weighted and therefore remains more sensitive to the influence of outlying scores. The item fit statistics are reported as the MnSq and Standardized z values (Zstd). For the purpose of the current study, infit and outfit MnSq should range from 0.6 to 1.4, and Zstd should range from −2 to +2 if items measure a single underlying unidimensional latent trait (Bond and Fox 2007; see also Linacre 2012). According to Table 6.2, the infit MnSq ranged from 0.83 to 1.24 and the outfit MnSq ranged from 0.79 to 1.20, all within the acceptable range of 0.6–1.4. Furthermore, the Infit Zstd ranged from −1.5 to 1.90, whereas the Outfit Zstd ranged from −1.90 to 1.60, all within the range of −2 to +2. Therefore, it can be concluded that the items fit the Rasch model sufficiently well to define a common underlying construct for the purpose of investigating VET attitudes at the group level.

The item measures in Table 6.2 indicate the endorsability of the items. The lower the measure, the easier the endorsement, and vice versa. In other words, items with lower measures are easier to be agreed with by test takers, whereas items with higher measure are more difficult to be agreed with. Item 1, 17, 2, and 3 (in ascending order of item measure) had the lowest measures, among which three items (1, 2, and 3) were in the content area of "test design" and one item (17) was in the area of "test-taking motivation." In addition, Items 13, 14, and 15, all in the area of "perceived interest of the VET tasks," had identical and very low measures (−0.15). Given that all items in this questionnaire were positively worded, the results indicate that test takers commented very positively on the design of the VET, including test delivery (Item 1), overall design of the test (Item 2), and the tasks in the VET (Item 3). On the whole, test takers had a strong internal motivation to take the VET as evidenced by their response to Item 17 (−0.86 logits). Compared to the tasks of "Reading" and "Repeat" in the VET (Item 11 and 12), respondents seemed to prefer the more constructed-response format, including "Questions" (Item 13), "Sentence Builds" (Item 14), and "Open Questions" (Item 15).

Conversely, the five items with the highest measures were Item 21, 5, 20, 19, and 8, among which three were in the content area of "test-taking motivation" (Items 19, 20, and 21) and two in the area of "perceived difficulty of the tasks in the VET" (Items 5 and 8). It is worth noting that the three motivation items, i.e., Item 19, 20, and 21, were all about the external motivation for taking the test such as the external recognition of the VET as valid proof of English proficiency (Item 19), obtaining the VET certificate (Item 20), and seeking employment (Item 21). The results suggested that test takers were more internally than externally motivated to take the VET. This finding came as no surprise to us for two reasons. First of all, all respondents participated in this study on a voluntary basis, and thus it is understandable that they could be more internally motivated to take the VET. Second, the VET is currently not widely known to Chinese university students, which could explain the relatively low external motivation. In addition, we also found that test takers did not think that the VET was a difficult test, as indicated by the measures of the items in the area of "perceived difficulty of the VET" (see Table 6.2). A possible explanation might be again attributed to the sample of this study. Since all respondents participated voluntarily, it is likely that they had high motivation in learning English and hence were more proficient in using English.

6.6.3 Reliability Estimates and Separation Indexes

In Rasch analysis, high item reliability estimate indicates a spread of items in the questionnaire from more difficult to easier, whereas high person reliability estimate indicates that the questionnaire administration spreads person scores from higher to lower. Acceptable item and person reliability estimates should be above the threshold of 0.8 (Bond and Fox 2007; Oon and Subramaniam 2011). In addition to reliability estimates, the Rasch measurement model also provides item and person separation indexes, which indicate the spread of item and person estimates along the measured variable in relation to the precision of those estimates (Bond and Fox 2007). The commonly accepted criterion for the separation indexes is at least 2.0 (Oon and Subramaniam 2011).

Item and person reliability estimates were 0.95 and 0.80, respectively, both could be considered as acceptably high. The item separation index was 4.60 (SD = 0.42) which is well above the acceptable criterion, whereas the person separation index was 1.97 (SD = 0.48) which is marginally below the acceptable threshold. The low person separation index might be attributed to the fact that the instrument was not sensitive enough to distinguish between high and low performers, and hence more items may be needed to improve individual person estimates (Boone et al. 2014). However, this is not the purpose of our survey, since we aimed at low-stakes group description of attitude.

6.6.4 Utility of Response Categories

Following the procedures suggested by Bond and Fox (2007), we investigated the utility of the response categories in the questionnaire. Specifically, Linacre's (2004) criteria were applied to verify the functioning of each response category which included (1) a minimum of 10 observations is needed for each category, (2) average category measures must increase monotonically with categories, (3) outfit MnSq 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 the probability curves should peak for each category (see also Oon and Subramaniam 2011, p. 125). Summary of category structure of the 6-point scale is presented in Table 6.3.

Results in Columns 2 and 3 demonstrate that except Category 1 (i.e., Strongly Disagree), the other five categories all had over 10 observations at the item level (i.e. >210), indicating that Category 1 was underused and might be redundant. The average measures increased monotonically from category 1 to 6 (i.e. −0.50–1.03), suggesting that these categories were used as expected by participants. Outfit MnSq statistics in Column 4 ranged from 0.93 to 1.09 (<2), suggesting that these categories did not introduce noise into the measurement process. The category probability curves (see Fig. 6.2) showed that each category emerged as a peak, although Categories 2, 3, and 4 showed low peaks. An examination of the distance between these three categories demonstrated that the distance between Category 2 (Disagree) and Category 3 (Slightly Disagree) was only 0.30 logits and that between Category 3 (Slightly Disagree) and Category 4 (Slightly Agree) was 0.74 logits, both failing to meet the acceptable range of 1.4–5. The results suggest that these three categories may not define distinct positions on the variable. A sensible solution in such circumstances, according to Bond and Fox (2007), is to collapse rating scale categories.

Table 6.3 Summary of category structure of the 6-point rating scale

Category	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1. Strongly disagree	90 (3)	−0.50	1.09	None
2. Disagree	220 (8)	−0.16	1.09	−1.25
3. Slightly disagree	532 (30)	0.02	0.85	−0.95
4. Slightly agree	806 (31)	0.34	0.93	−0.21
5. Agree	700 (27)	0.69	0.93	0.64
6. Strongly agree	277 (11)	1.03	1.07	1.78

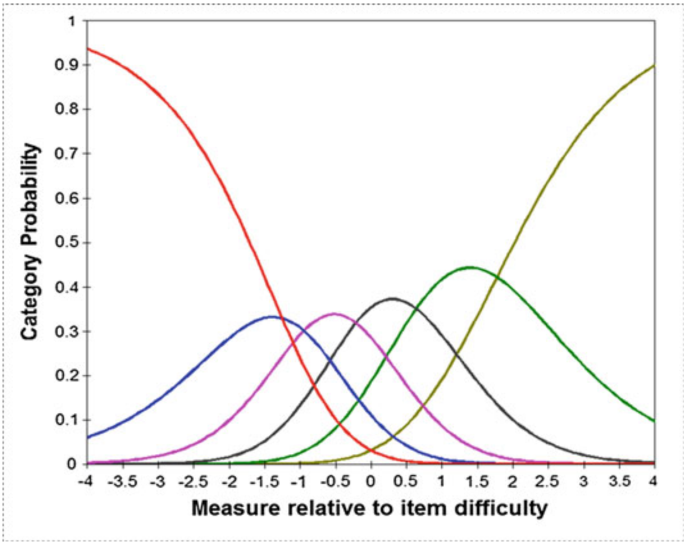
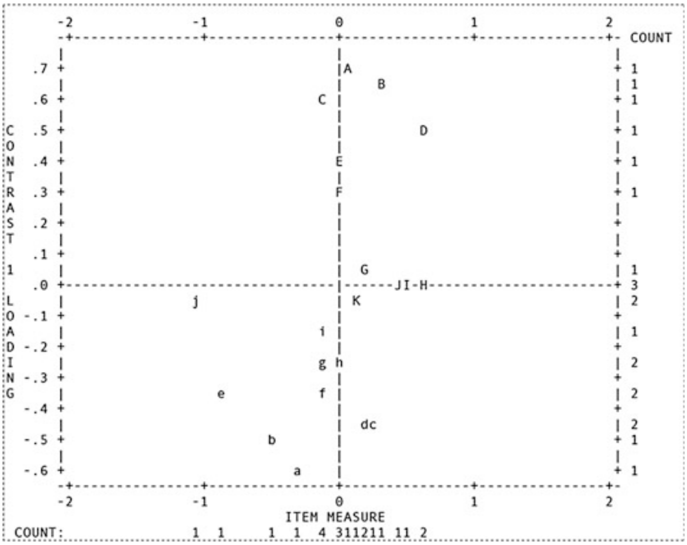


Fig. 6.2 Category probability curves for the 6-point scale

Given that the distance between Category 2 and Category 3 was only 0.30 logits, an attempt was made to collapse these two categories, and hence the rating scale was reorganized from 123456 to 122345. However, collapsing these two categories did not improve the category threshold between Category 2 and Category 3 and between Category 3 and Category 4, and item and person reliability remained 0.95 and 0.80, respectively. This same issue was reported by Oon and Subramaniam (2011) in their investigation of the category structure of a six-point scale. The reason might reside in the fact that some categories were redundant in the rating scale which could not define distinct positions on the variable. Although the optimal number of response categories used in a scale remains contentious (Bond and Fox 2007), findings derived from this present study suggest that a scale using a smaller number of categories can be attempted in future investigation of test takers' attitude toward the VET.

6.6.5 *Rasch Factor Analysis*

Linacre (2012) further suggested using Rasch Factor Analysis of the item-person residuals to investigate data dimensionality. According to Bond and Fox (2007, p. 253), the term Rasch Factor Analysis is somewhat misleading which, as a matter of fact, involves first a regular Rasch analysis procedure, followed by a factor analysis of the residuals that remain after the linear Rasch measure has been



```
      -2              -1               0               1               2
      -+---------------+---------------+---------------+---------------+- COUNT
      |                                |                                |
   .7 +                                |A                               + 1
      |                                |    B                           | 1
   .6 +                              C |                                + 1
      |                                |                                |
C  .5 +                                |         D                      + 1
O     |                                |                                |
N  .4 +                                E                                + 1
T     |                                |                                |
R  .3 +                                F                                + 1
A     |                                |                                |
S  .2 +                                |                                +
T     |                                |                                |
   .1 +                                |                                +
1     |                                |  G                             | 1
   .0 +--------------------------------|------JI-H----------------------+ 3
L     |               j                | K                              | 2
O -.1 +                                |                                +
A     |                              i |                                | 1
D -.2 +                                |                                +
I     |                              g h                                | 2
N -.3 +                                |                                +
G     |                  e           f |                                | 2
  -.4 +                                |                                +
      |                                |  dc                            | 2
  -.5 +                        b       |                                + 1
      |                                |                                |
  -.6 +                           a    |                                + 1
      -+---------------+---------------+---------------+---------------+-
      -2              -1               0               1               2
                                 ITEM MEASURE
 COUNT:               1  1     1  1  4 311211 11 2
```

Fig. 6.3 Plot of standardized residuals of the scale

extracted from the data set. In this study, Principal Component Analysis was used to investigate whether a subdimension existed in the variance unexplained or unmodeled by the primary Rasch measure. The results are illustrated in Fig. 6.3. Linacre (2012) suggested that variance greater than or equal to 50 % for the Rasch dimension can be regarded as support for the claim that the scale is unidimensional. If a significant amount of variance is found in the second dimension (first contrast), the scale might contain competing dimensions. On the other hand, if the first contrast has the strength of less than 3 items and the unexplained variance by the first contrast is less than 5 %, then the unidimensionality of the scale is supported.

In this analysis, the Rasch dimension explained 37.5 % of the variance, below the desired (50 %). The second dimension had an eigenvalue of 3.2 and accounted for 10.5 % of the variance that was not modeled. As illustrated in Fig. 6.3, a distinct cluster of Items A, B, and C (corresponding to Item 3, 2, and 12, respectively) showed the largest contrast loadings (>0.50), and they therefore might comprise an additional dimension. Inspection of the three items did not reveal a meaningful subdimension. At the stage of questionnaire development, these three items were intended to reflect the theoretical framework of attitude which guided this investigation. In addition, these items went through some a priori validation procedures such as expert judgment. Given that the Rasch measure explained 37.5 % of the variance which was not far below the criterion of 50 % and the eigenvalue of the second dimension (3.2), it is reasonable to assume that this scale is sufficiently unidimensional for our purposes and tapped into one underlying construct.

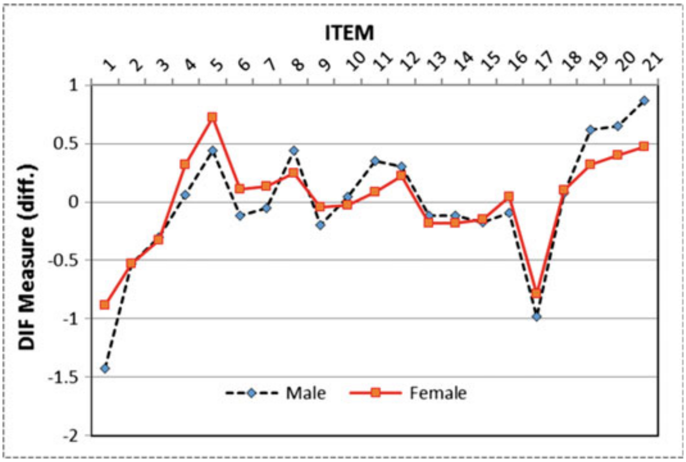


Fig. 6.4 Plot of item estimates between male and female test takers

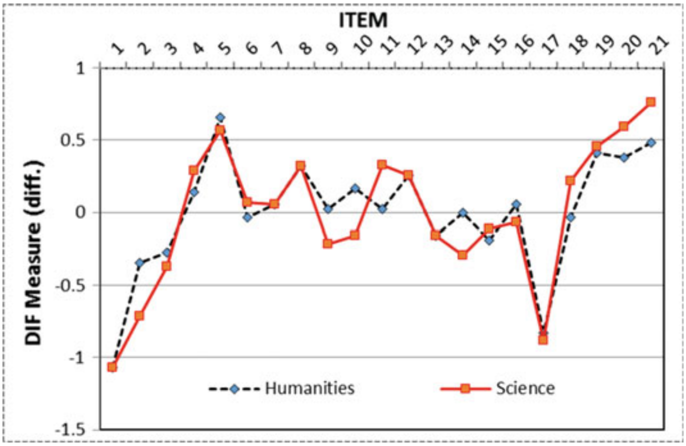


Fig. 6.5 Plot of item estimates between test takers from different backgrounds

6.6.6 *Differential Item Functioning*

Measurement invariance is a crucial property of scientific measurement. The principle of measurement invariance requires that the difficulty in the items should remain stable across the two different subsamples of interest (Bond and Fox 2007). To investigate whether the invariance principle held, Rasch measures for female and male test takers as well as test takers from different backgrounds (i.e. humanities and science) were calculated, and the results are illustrated in Figs. 6.4 and 6.5.

As revealed by these two figures, some items appeared slightly more difficult for one subsample than the other to endorse while some items functioned conversely. However, whether the DIF is substantive depends on the DIF contrast which is the difference in item measures between the subsamples. The contrast should be at least 0.50 logits for DIF to have an impact. In addition, *t*-test result can aid in the interpretation of the magnitude of the DIF contrast. A significant *t*-test indicates noticeable DIF (Linacre 2012). An inspection of the DIF contrast reveals that for male and female test takers, only one item (Item 1) was found to display a DIF contrast of −0.54, marginally above the criterion of 0.50. This item was more difficult for females (−0.88 logits) than for males (−1.42 logits) to endorse. The *t*-test result indicated that the difference was a statistically significant one (t = −2.36, df = 104, $p < 0.05$). With regard to DIF analysis, Linacre (2012) cautioned that DIF impact on the person measures could be affected by the length of the test. The longer a test is, the less likely that a DIF size is statistically significant. Given that the scale used in this study was relatively short with 21 items and the DIF contrast was only −0.54 logits, we could assume that this item functioned sufficiently equally between male and female test takers. This conclusion was further corroborated by the DIF analysis performed on test takers from the two academic backgrounds, i.e.,. humanities and science, as revealed by Fig. 6.5. No items were found to display a DIF contrast of over 0.50 logits. These findings could therefore quite safely bring us to the conclusion that the principle of measurement invariance held for the scale.

6.7 Conclusions

Despite the increasingly extensive application of Rasch modeling in human sciences (e.g., Bond and Fox 2007; Cavanagh and Waugh 2011), the use of Rasch measurement in the field language testing remains limited and is largely confined to the investigation of test validity (McNamara and Knoch 2012). Few attempts have been made to use the Rasch measurement models to analyze Likert-type scale survey data. To overcome the limitations of the CTT in data analysis, in this study we used the RSM in Rasch measurement to investigate test takers' attitude toward the VET, a fully automated spoken English test. A 6-point Likert-style attitude questionnaire was developed on the basis of the theoretical conceptualization of attitude which was articulated by Murray et al. (2012) and mirrored the understanding of this construct in the realm of social psychology. All collected data were then subjected to Rasch analysis.

On the whole, the data fit the Rasch model well with reasonably high item and person reliability estimates. The satisfactory data-model fit was also evidenced by the Rasch-modeled item statistics, including PTMEA correlations, infit and outfit MnSq, and Zstd values. Investigation of the category structure demonstrated that some response categories in the scale failed to define distinct positions on the latent variable, suggesting that the number of response categories could be reduced in future investigations. Rasch factor analysis suggested the existence of a small secondary dimension although inspection of the items did not reveal a meaningful interpretation. DIF analysis indicated that the principle of measurement invariance held for the subsamples defined by both gender and academic background. These findings bring us to the conclusion that the scale is essentially unidimensional and measured test takers' attitude toward the VET although refinements are warranted to further improve the utility and validity of the scale such as adding more difficult items and reducing the number of categories.

In addition to the aforementioned findings about the rating scale, Rasch-modeled item measures indicate that test takers generally enjoyed their experience of taking the VET, and in particular, they commented very positively on the design of the VET. Needless to say, these findings should be encouraging to the VET provider who is keen to improve and promote the use of the test in China and elsewhere. Compared to the selected-response tasks in the VET, test takers preferred the constructed-response formats such as Story Retelling and Open Questions. Not surprisingly, these volunteer test takers were found to be more internally than externally motivated to take the VET. The findings of this study largely concur with Chun's (2006) review of the Versant suite of tests, suggesting that one potential problem besetting the VET lies in the apparent lack of task authenticity. Test takers prefer the constructed-response tasks because they claim that such tasks are more authentic and can therefore better reflect their spoken English ability (see also Fan 2014). How to improve authenticity without sacrificing the scientific rigor in an automated spoken English test remains a key future challenge facing the VET provider and other automated spoken language test providers.

6.8 Limitations and Implications

Because of practical and logistical constraints, this study adopted convenience sampling, and consequently, all participants in this study were university students who participated on a voluntary basis. It should be noted, however, that the VET is not targeted exclusively at university students. In fact, it is designed for a wide range of potential test takers, including students, staff, or officers working with academic institutions, corporations, and government agencies (Pearson 2008, p. 3). In addition, given that all participants in this study were volunteers, caution needs to be exercised in the interpretations of relevant research findings, in particular in terms of test-taking motivation.

Second, test anxiety has been included in some previous attitudinal investigations (e.g., Zhao and Cheng 2010) but was not investigated in this study due to the low-stakes nature of this test for this particular sample. It would be worthwhile to investigate this attitudinal aspect in future research if the VET is used for high-stakes purposes such as admission, selection, or recruitment.

Finally, the Rasch analysis we performed on the collected data, albeit quite encouraging, also exposed a few problems with the quality of the attitude questionnaire such as low person separation index, lack of items targeted at persons with high ability, and some response categories failing to define distinct positions on the variable. The scale warrants further revision and validation so as to ensure that it is unidimensional and tapped into the underlying construct, i.e., test takers' attitude toward the VET.

This study has implications for both the VET provider and other researchers working in the field of language testing. For the VET provider, this study presents some credible evidence as to how test takers viewed this automated spoken English test, thus paving the way for the future promotion and improvement of this test. Taking into consideration the important role of test takers' attitude in test validation (e.g., Alderson et al. 1995; Messick 1989), the VET provider should properly address test takers' concerns emerging from this investigation. For example, the VET provider might consider how to make this test more authentic and introduce tasks which better reflect real-life interactive language use in the test. Furthermore, this study has methodological implications for researchers working in the field of language testing. At present, the vast majority of the survey data in language testing have been subjected to various CTT analyses such as correlational and factor analyses. The validity of these analyses is often, if not always, questionable because of the inherent limitations of the CTT in analyzing Likert-type scale data. This study therefore demonstrates how the Rasch measurement models can be used to effectively investigate the quality of rating scales and to yield reliable and valid research results from analyzing Likert-type scale attitudinal data.

Acknowledgments The preparation of this manuscript was supported by the National Social Sciences Fund of the People's Republic of China under the project title "the Development and Validation of Standards in Language Testing" (Grant Nunber: 13CYY032) and a joint research project between the Language Testing Center, Fudan University and Pearson Knowledge Technologies. This study used the same data as reported in Fan (2014). We would like to thank Prof. Peiying Ji, Prof. Jiangbo Wan, and Ms. Diana Hon for their assistance in collecting the data for this study. Also, we would like to thank all the students who participated in this study.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Allport, G. W. (1971). Attitudes. In K. Thomas (Ed.), *Attitudes and behavior*. Harmondsworth, UK: Penguin.

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42.

Baker, C. (1988). *Key issues in bilingualism and bilingual education*. Clevedon, The UK: Multilingual Matters.

Baker, C. (1992). *Attitudes and language*. Clevedon, The UK: Multilingual Matters.

Bernstein, J., Van-Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377.

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

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York and London: Springer, Dordrecht Heidelberg.

Brown, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277–301. doi:10.1177/026553229301000305.

Brown, D. H. (2000). *Principles of language learning and teaching* (4th ed.). New York: Longman.

Cavanagh, R. F., & Waugh, R. F. (Eds.). (2011). *Applications of rasch measurement in learning environments research*. Rotterdam; Boston: Sense Publishers.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly, 3*(3), 295–306.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. New York: Handcourt Brace Jovanovich.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test taker have to offer? *Language Testing, 19*(4), 347–368.

Embreston, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey, NJ: Lawrence Erlbaum.

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

Farhady, H. (2008). *Human operated, machine mediated, and automated tests of spoken English*. Paper presented at the American Association of Applied Linguistics (AAAL) Conference, Washington, D. C.

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

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.

Ingram, E. (1977). Basic concepts in testing. In J. P. B. Allen & A. Davies (Eds.), *Edinburgh course of applied linguistics* (Vol. 4). Oxford: Oxford University Press.

Jin, Y., & Cheng, L. (2013). The effects of psychological factors on the validity of high-stakes tests. *Modern Foreign Languages, 36*(1), 62–69.

Karelitz, T. M. (2014). Using public opinion to inform the validation of test scores (Research report). Retrieved Novermber 17, 2014, from www.nite.org.il/files/reports/e387.pdf.

Ladegaard, H. J. (2000). Language attitudes and sociolinguistic behavior: Exploring attitude-behavior relations in language. *Journal of Sociolinguistics, 4*(2), 214–233.

Linacre, J. M. (2004). Optimal rating scale category effectiveness. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press.

Linacre, M. (2012). Winsteps Tutorial. Retrieved November 7, 2014, from http://www.winsteps.com/tutorials.htm.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). McMillan: American Council on Education.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 553–574.

Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW Australia. *Language Testing, 29*(5), 577–595.

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.

Pearson. (2008). Versant English test: Test description and validation summary. Retrieved November 7, 2014, from www.versanttest.co.uk/pdf/ValidationReport.pdf.

Present-Thomas, R., & Van-Moere, A. (2009). *NRS classification consistency of two spoken English tests*. Paper presented at the East Coast Organization of Language Testers (ECOLT) Conference, Washington, D. C.

Rasch, G. (1960). *Probalistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Paedagogiske Institut.

Rasch, G. (1980). *Probalistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Rasti, I. (2009). Iranian candidates' attitudes towards the IELTS. *Asian EFL Journal, 11*(3), 110–155.

Reid, N. (2006). Thoughts on attitude measurement. *Research in Science and Technological Education, 24*(1), 3–27.

Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Pearson Education.

Stevenson, D. K. (1985). Authenticity, validity, and a tea party. *Language Testing, 2*(1), 41–47.

Wu, J. (2008). Views of Taiwanese students and teachers on English language testing. *University of Cambridge ESOL Examinations Research Note, 34*(2), 6–9.

Zhao, J., & Cheng, L. (2010). Exploring the relationship between Chinese University students' attitude towards the College English Test and their test performance. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese leaner* (pp. 190–218). New York and London: Routledge, Taylor & Francis Group.

Chapter 7
The Application of Rasch in the Validation of Corporate Citizenship Scale

Kamala Vainy Pillai, Rajah Rasiah and Geoffrey Williams

Abstract This paper articulates the application of Rasch measurement in corporate citizenship research. With burgeoning expectation for greater corporate responsibility, studies have found that many companies continue to resort to green washing tactics in order to cope with growing pressures. In the absence of systematic adoption, the concept of corporate citizenship may remain rhetoric. The study is aimed at determining the fundamental attributes that facilitate the internalisation of corporate citizenship within companies. The study had to address two main challenges: First, the small sample size expected as it was based on primary data collection explicitly seeking the views of managers practicing corporate citizenship and second, as the study applied a multidisciplinary approach, there was a lack of prior systematic research available. Rasch modelling was applied to establish a psychometrically sound scale for the purpose of this study. A pilot test was followed with a larger sample set, where a total of 634 companies listed on the Malaysian Exchange were surveyed through online survey, of which 100 companies responded. The instrument's reliability and validity were conducted using Winsteps version 3.49. The results of Rasch modelling analysis indicated the items measured reliability (r = 0.93) and persons measured reliability (r = 0.97). In addition, both item separation (3.51) and person separation (5.27) were found to be statistically significant. Further, all items measured in the same direction (point-measure correlation > 0.30) and valid items showed good item fit and

K.V. Pillai (✉)
Faculty of Business and Humanities, Curtin University, Miri, Malaysia
e-mail: kamala.pillai@curtin.edu.my; kamalavarny@yahoo.com

R. Rasiah
Department of Development Studies, Faculty of Economics and Administration, University of Malaya, Kuala lumpur, Malaysia
e-mail: rajah@um.edu.my

G. Williams
Education Management, University Tun Abdul Razak, Kuala lumpur, Malaysia
e-mail: geoffreyalanwilliams@gmail.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_7

constructed a continuum of increasing intensity. This study's significance stands on its contribution towards the application of Rasch by future researchers in diverse social science and industrial settings in developing and validating sound scales.

Keywords Rasch model · Corporate citizenship · Sustainability · Public listed companies

7.1 Introduction

Although the widespread implementation of corporate citizenship standards and frameworks globally is emblematic of a positive trajectory towards sustainability practises, ongoing studies (Pillai 2013; Castka et al. 2004; Frankental 2001; Olsen 2002; Mcintosh et al. 2003) continue to reveal the lack of internalising capability of these frameworks. It is evident that while these standards and frameworks facilitated adoption of sustainability practises for compliance and audit, they offer limited knowledge on how a sustainable business should be run (Bamber et al. 2000; Castka et al. 2004). This state of affair has inadvertently resulted in a common scenario, where sustainability practises continue to be implemented as stand-alone systems in business. Such stand-alone systems not only impede internal control and embedding of sustainability conscience but it also reaffirms the criticism against corporations on their sustainability reports as mere public relations gimmicks. This exposition juxtaposed with the heightening corporate scandals and misdemeanours by seemingly responsible companies such as Coca Cola, Xerox, Olympus, Satyam, Tyco, WorldCom, Nike and Sino-Forest Corp elucidate a missing link in the corporate citizenship body of knowledge.

In determining the missing link, the research leveraged insights from a multi-disciplinary study of learning organisation and organisational development body of knowledge. Here, a review of literature brought forth the articulation of 'hardware' and 'software' (or also known as soft) elements. While standards and frameworks provided the hardware, the presence of soft elements within an organisation was needed to facilitate intrinsic change in corporate conduct. The significance of 'software' or intrinsic elements was often underestimated (Antony and Bhattacharyya 2010; Jung and Hong 2008; Hermel and Remis-Pujol 2003) and thus justified further examination.

In the context of this study, the research adopted the concurrent mixed methods procedure as part of a Mixed Methods strategy. This procedure involves the convergence of both quantitative and qualitative data in order to derive a comprehensive analysis of the research problem. This paper expounds specifically on the application of Rasch measurement model in establishing the validation of a psychometrically sound scale for the purpose of this study.

As the application of Rasch measurement model is considerably new in corporate citizenship research. The following section sets the background and justification for the adoption of this measurement model.

7.2 Application of Rasch Measurement Model

The shift from the traditional measurement theory in social science based on classical test theory or true-score model (TSM) to latent-trait models became strongly evident after the 1960s (Searing 2008; Smith 1999), attributed to the plaguing deficiencies in TSM in terms of the vulnerability of the scale of measurement to respondents measured. Here, it is pertinent to acknowledge that in social science research, various statistical studies require interval-level data. Measures undertaken, however, do not aptly demonstrate that data meet the requirements of interval scale measure. The argument is that statistical testing on scales that are, in fact, ordinal in nature and represented as interval is based on faulty assumptions. Rasch modelling addresses this contention by facilitating the endorsement of interval scale from raw data through the measurement of distance between individual items expressed in terms of log odds unit, also known as logits (Searing 2008), which eliminate the risk of tabulation and assumptions based on ordinate scale. In addition, Rasch model demonstrates objective measurement in the requirement for unidimensionality. This simply means that in order to meet the criteria for unidimensionality, only the construct being measured should be reflected in the scale. Unidimensionality is measured in terms of local independence and specific objectivity.

The most prominent feature of Rasch is that it is a theory-driven process. This means that the theoretical model under measure should be independent of the object measured. Hence, the data are examined for their fit to the theoretical model under study. In Rasch, this is achieved through the separation of item and person parameters measurement (Smith 1999). In sum, data fit to the model facilitates the determination of internal construct validity of the measure, which includes the ordering of categories, unidimensionality and whether or not items work in the same way across different groups.

Since its inception, Rasch analysis has been applied in the field of psychology, health, education and social science (Andrich 1988, 1985, 1982; Duncan 1984) to assess the psychometric properties of a scale. Psychometrics refers to the design and interpretation of quantitative tests that measures psychological variables such as aptitude, behaviour, intelligence and personality traits. The application of Rasch in social science studies includes testing intrinsic motivation on satisfaction of the wireless Internet service (Islam 2011), drinking behaviour (Mcintosh et al. 2006), domestic food safety practises (Fischer et al. 2006) and measuring psychological distress (Pallant and Tennant 2007) based on primary data collection.

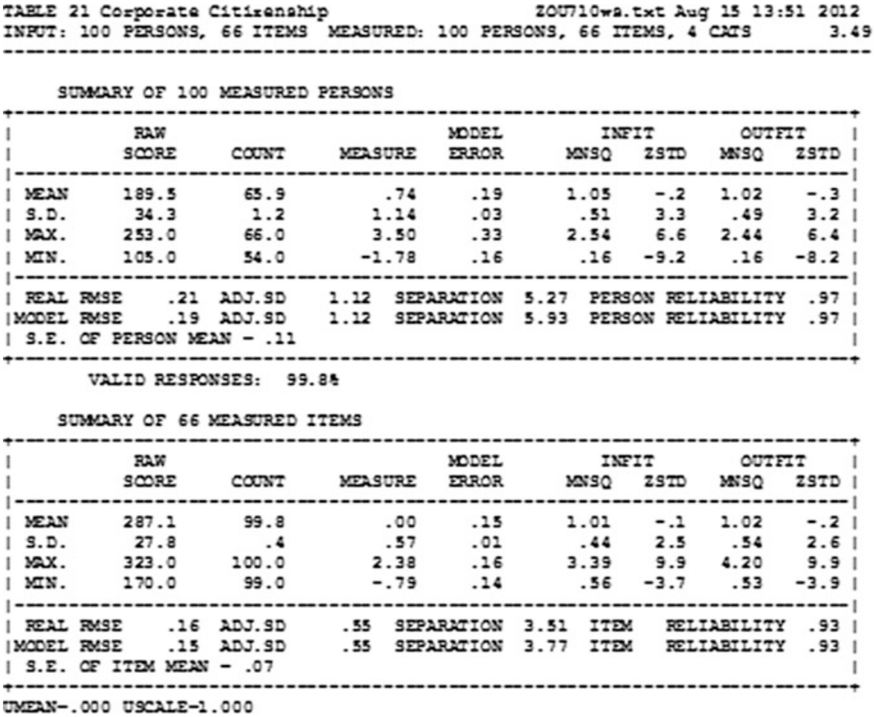
In the following section, the diagnostics applied in this study using Winsteps version 3.49 for Rasch analysis are deliberated at length.

7.3 Rasch Analysis and Findings

Under Rasch analysis, several criteria must be met by a model in order to establish validity. Hence, using Winsteps version 3.49, the diagnostics applied included (i) summary statistics, (ii) item fit order, (iii) item polarity map, (iv) item-person map and (v) principal component.

Table 7.1 represents the summary statistics of the 66 items generated from Rasch analysis. The items measured reliability ($r = 0.93$) and the persons measured reliability ($r = 0.97$). Both item separation (3.51) and person separation (5.27) were found to be statistically significant. According to Bond and Fox (2010), item reliability index of 0.93 is high and indicates confidence in the replicability of item placement across other samples. In addition, the high person reliability lends further support that the study was able to expect consistency of inferences. Items mean was 287.1 and standard deviation (SD) 27.8; similarly, the persons M was 189.5 and SD was 34.3.

Table 7.1 Summary statistics



```
TABLE 21 Corporate Citizenship                      ZOU710wa.txt Aug 15 13:51 2012
INPUT: 100 PERSONS, 66 ITEMS  MEASURED: 100 PERSONS, 66 ITEMS, 4 CATS         3.49
----------------------------------------------------------------------------------

     SUMMARY OF 100 MEASURED PERSONS
+-------------------------------------------------------------------------------+
|           RAW                             MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE     ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN      189.5       65.9         .74      .19       1.05    -.2   1.02    -.3 |
| S.D.       34.3        1.2        1.14      .03        .51    3.3    .49    3.2 |
| MAX.      253.0       66.0        3.50      .33       2.54    6.6   2.44    6.4 |
| MIN.      105.0       54.0       -1.78      .16        .16   -9.2    .16   -8.2 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .21  ADJ.SD    1.12  SEPARATION  5.27  PERSON RELIABILITY  .97 |
|MODEL RMSE    .19  ADJ.SD    1.12  SEPARATION  5.93  PERSON RELIABILITY  .97 |
| S.E. OF PERSON MEAN - .11                                                     |
+-------------------------------------------------------------------------------+
        VALID RESPONSES:   99.8%

     SUMMARY OF 66 MEASURED ITEMS
+-------------------------------------------------------------------------------+
|           RAW                             MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE     ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN      287.1       99.8         .00      .15       1.01    -.1   1.02    -.2 |
| S.D.       27.8         .4         .57      .01        .44    2.5    .54    2.6 |
| MAX.      323.0      100.0        2.38      .16       3.39    9.9   4.20    9.9 |
| MIN.      170.0       99.0        -.79      .14        .56   -3.7    .53   -3.9 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .16  ADJ.SD     .55  SEPARATION  3.51  ITEM   RELIABILITY  .93 |
|MODEL RMSE    .15  ADJ.SD     .55  SEPARATION  3.77  ITEM   RELIABILITY  .93 |
| S.E. OF ITEM MEAN - .07                                                       |
+-------------------------------------------------------------------------------+
UMEAN-.000 USCALE-1.000
```

7.4 Item Polarity Map

The item polarity map, shown in Table 7.2, illustrates that all valid items measured in the same direction (point-measure correlation > 0.40), which is greater than the recommended value (0.30).

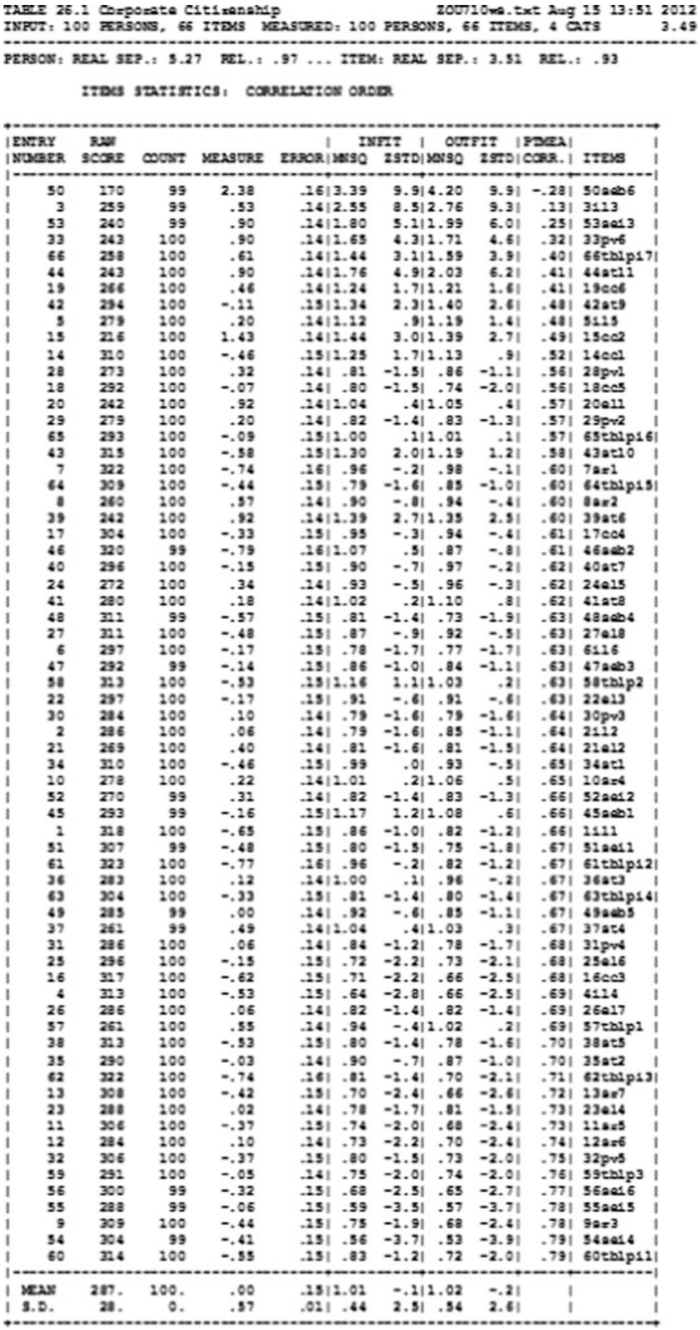
7.5 Item Map

The item-person map is also known as the Wright Map (Bond and Fox 2010). According to Smith (1993), a unique strength of the Rasch model is its requirement that the outcome of any interaction between person and item to be solely determined by just two parameters, that is, the ability of the person and the difficulty in the item. This requirement, therefore, establishes a strong framework to test data for the presence of anomalous behaviour that may influence the estimation of item and person parameters. Bond and Fox (2010) reiterated that the identification of anomalies in Rasch Modelling was not restricted to guessing and addressed any potential measurement disturbance. For example, if random guessing occurred, Rasch Modelling enabled the study to detect measurement disturbance. Here, Table 7.3 plots the item difficulty estimates for all 66 items relative to the person distribution along the logit scale. The map exhibits that the valid items are located in their calibrations on a single continuum. As shown on the right side of Table 7.3, the items on this single continuum are structured from the "less difficult corporate citizenship excellence items" to the "more difficult corporate citizenship excellence items." The left side of this map also matches the level of a person's ability in a single line, where respondents are ordered from the "higher level of ability to endorse public limited companies' corporate conduct with the corporate citizenship excellence" to the "lower level agreement to endorse public limited companies' corporate conduct with the corporate citizenship excellence." In sum, as shown in Table 7.3, the analysis output demonstrates that the most difficult item was 50seb6 and the least difficult 46seb2 and that the majority of senior managers of the public listed companies elected on the performance of corporate citizenship.

7.6 Item Fit Order

Here, the fit indices help the study to ascertain whether the assumption of unidimensionality holds up empirically (Bond and Fox 2010). Rasch analysis programs usually report fit statistics as two chi-square ratios, infit and outfit mean square (MNSQ) statistics (Wright 1984; Wright and Masters 1982), to determine the compatibility of the empirical data with the requirements of the model. As shown in Table 7.4, the item fit order indicates that most items show good item fit and

Table 7.2 Item polarity map

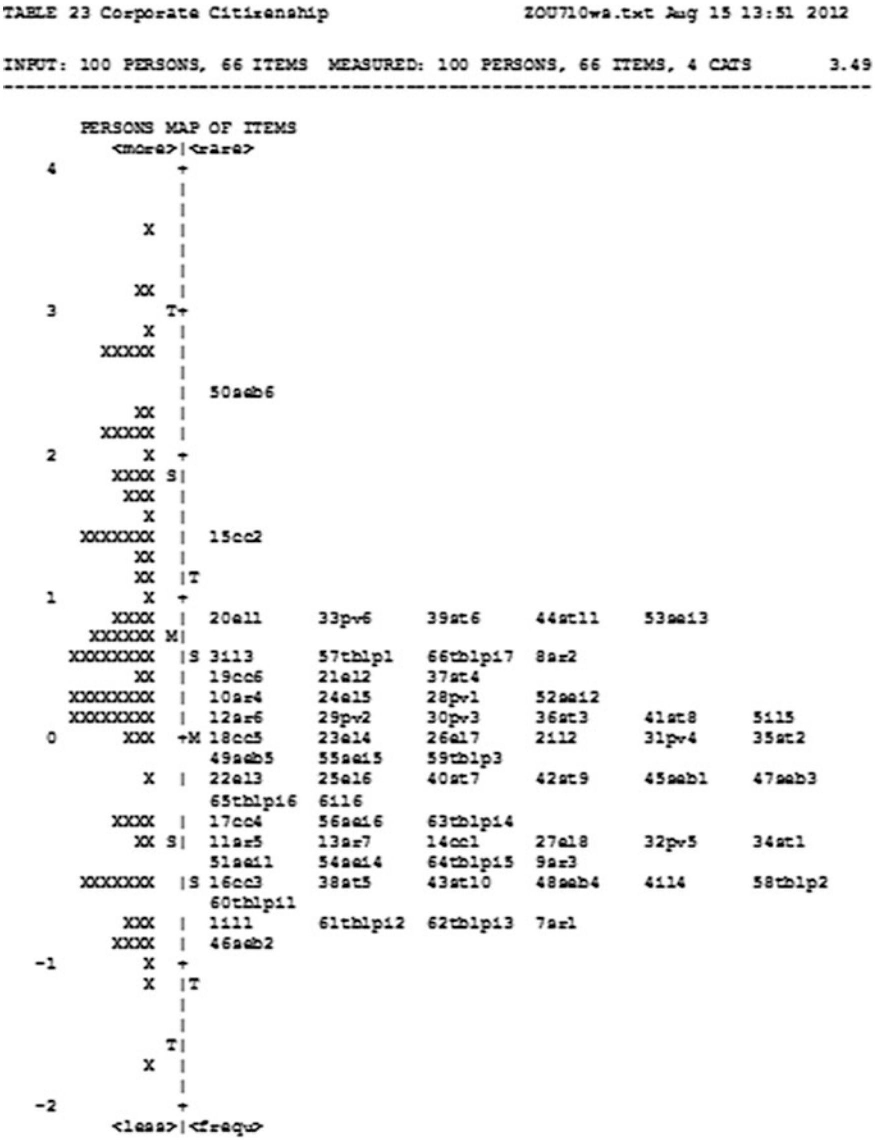


```
TABLE 26.1 Corporate Citizenship              ZOU710wa.txt Aug 15 13:51 2012
INPUT: 100 PERSONS, 66 ITEMS  MEASURED: 100 PERSONS, 66 ITEMS, 4 CATS       3.49
--------------------------------------------------------------------------------
PERSON: REAL SEP.: 5.27  REL.: .97 ... ITEM: REAL SEP.: 3.51  REL.: .93

        ITEMS STATISTICS:  CORRELATION ORDER

+---------------------------------------------------------------------------------+
|ENTRY    RAW                              |  INFIT   |  OUTFIT  |PTMEA|          |
|NUMBER   SCORE   COUNT  MEASURE      ERROR|MNSQ  ZSTD|MNSQ  ZSTD|CORR.| ITEMS    |
|------------------------------------------+----------+----------+-----+----------|
|      50    170     99     2.38        .16|3.39   9.9|4.20   9.9| -.28| 50aeb6   |
|       3    259     99      .53        .14|2.55   8.5|2.76   9.3|  .13| 3i13     |
|      53    240     99      .90        .14|1.80   5.1|1.99   6.0|  .25| 53aei3   |
|      33    243    100      .90        .14|1.65   4.3|1.71   4.6|  .32| 33pv6    |
|      66    258    100      .61        .14|1.44   3.1|1.59   3.9|  .40| 66tblpi17|
|      44    243    100      .90        .14|1.76   4.9|2.03   6.2|  .41| 44at11   |
|      19    266    100      .46        .14|1.24   1.7|1.21   1.6|  .41| 19cc6    |
|      42    294    100     -.11        .15|1.34   2.3|1.40   2.6|  .48| 42at9    |
|       5    279    100      .20        .14|1.12    .9|1.19   1.4|  .48| 5i15     |
|      15    216    100     1.43        .14|1.44   3.0|1.39   2.7|  .49| 15cc2    |
|      14    310    100     -.46        .15|1.25   1.7|1.13    .9|  .52| 14cc1    |
|      28    273    100      .32        .14| .81  -1.5| .86  -1.1|  .56| 28pv1    |
|      18    292    100     -.07        .14| .80  -1.5| .74  -2.0|  .56| 18cc5    |
|      20    242    100      .92        .14|1.04    .4|1.05    .4|  .57| 20el1    |
|      29    279    100      .20        .14| .82  -1.4| .83  -1.3|  .57| 29pv2    |
|      65    293    100     -.09        .15|1.00    .1|1.01    .1|  .57| 65tblpi16|
|      43    315    100     -.58        .15|1.30   2.0|1.19   1.2|  .58| 43at10   |
|       7    322    100     -.74        .16| .96   -.2| .98   -.1|  .60| 7ar1     |
|      64    309    100     -.44        .15| .79  -1.6| .85  -1.0|  .60| 64tblpi15|
|       8    260    100      .57        .14| .90   -.8| .94   -.4|  .60| 8ar2     |
|      39    242    100      .92        .14|1.39   2.7|1.35   2.5|  .60| 39at6    |
|      17    304    100     -.33        .15| .95   -.3| .94   -.4|  .61| 17cc4    |
|      46    320     99     -.79        .16|1.07    .5| .87   -.8|  .61| 46aeb2   |
|      40    296    100     -.15        .15| .90   -.7| .97   -.2|  .62| 40at7    |
|      24    272    100      .34        .14| .93   -.5| .96   -.3|  .62| 24el5    |
|      41    280    100      .18        .14|1.02    .2|1.10    .8|  .62| 41at8    |
|      48    311     99     -.57        .15| .81  -1.4| .73  -1.9|  .63| 48aeb4   |
|      27    311    100     -.48        .15| .87   -.9| .92   -.5|  .63| 27el8    |
|       6    297    100     -.17        .15| .78  -1.7| .77  -1.7|  .63| 6i16     |
|      47    292     99     -.14        .15| .86  -1.0| .84  -1.1|  .63| 47aeb3   |
|      58    313    100     -.53        .15|1.16   1.1|1.03    .2|  .63| 58tblp2  |
|      22    297    100     -.17        .15| .91   -.6| .91   -.6|  .63| 22el3    |
|      30    284    100      .10        .14| .79  -1.6| .79  -1.6|  .64| 30pv3    |
|       2    286    100      .06        .14| .79  -1.6| .85  -1.1|  .64| 2i12     |
|      21    269    100      .40        .14| .81  -1.6| .81  -1.5|  .64| 21el2    |
|      34    310    100     -.46        .15| .99    .0| .93   -.5|  .65| 34at1    |
|      10    278    100      .22        .14|1.01    .2|1.06    .5|  .65| 10ar4    |
|      52    270     99      .31        .14| .82  -1.4| .83  -1.3|  .66| 52aei2   |
|      45    293     99     -.16        .15|1.17   1.2|1.08    .6|  .66| 45aeb1   |
|       1    318    100     -.65        .15| .86  -1.0| .82  -1.2|  .66| 1i11     |
|      51    307     99     -.48        .15| .80  -1.5| .75  -1.8|  .67| 51aei1   |
|      61    323    100     -.77        .16| .96   -.2| .82  -1.2|  .67| 61tblpi12|
|      36    283    100      .12        .14|1.00    .1| .96   -.2|  .67| 36at3    |
|      63    304    100     -.33        .15| .81  -1.4| .80  -1.4|  .67| 63tblpi14|
|      49    285     99      .00        .14| .92   -.6| .85  -1.1|  .67| 49aeb5   |
|      37    261     99      .49        .14|1.04    .4|1.03    .3|  .67| 37at4    |
|      31    286    100      .06        .14| .84  -1.2| .78  -1.7|  .68| 31pv4    |
|      25    296    100     -.15        .15| .72  -2.2| .73  -2.1|  .68| 25el6    |
|      16    317    100     -.62        .15| .71  -2.2| .66  -2.5|  .68| 16cc3    |
|       4    313    100     -.53        .15| .64  -2.8| .66  -2.5|  .69| 4i14     |
|      26    286    100      .06        .14| .82  -1.4| .82  -1.4|  .69| 26el7    |
|      57    261    100      .55        .14| .94   -.4|1.02    .2|  .69| 57tblp1  |
|      38    313    100     -.53        .15| .80  -1.4| .78  -1.6|  .70| 38at5    |
|      35    290    100     -.03        .14| .90   -.7| .87  -1.0|  .70| 35at2    |
|      62    322    100     -.74        .16| .81  -1.4| .70  -2.1|  .71| 62tblpi13|
|      13    308    100     -.42        .15| .70  -2.4| .66  -2.6|  .72| 13ar7    |
|      23    288    100      .02        .14| .78  -1.7| .81  -1.5|  .73| 23el4    |
|      11    306    100     -.37        .15| .74  -2.0| .68  -2.4|  .73| 11ar5    |
|      12    284    100      .10        .14| .73  -2.2| .70  -2.4|  .74| 12ar6    |
|      32    306    100     -.37        .15| .80  -1.5| .73  -2.0|  .75| 32pv5    |
|      59    291    100     -.05        .14| .75  -2.0| .74  -2.0|  .76| 59tblp3  |
|      56    300     99     -.32        .15| .68  -2.5| .65  -2.7|  .77| 56aei6   |
|      55    288     99     -.06        .15| .59  -3.5| .57  -3.7|  .78| 55aei5   |
|       9    309    100     -.44        .15| .75  -1.9| .68  -2.4|  .78| 9ar3     |
|      54    304     99     -.41        .15| .56  -3.7| .53  -3.9|  .79| 54aei4   |
|      60    314    100     -.55        .15| .83  -1.2| .72  -2.0|  .79| 60tblpi11|
|------------------------------------------+----------+----------+-----+----------|
| MEAN      287.   100.      .00        .15|1.01   -.1|1.02   -.2|     |          |
| S.D.       28.     0.      .57        .01| .44   2.5| .54   2.6|     |          |
+---------------------------------------------------------------------------------+
```

Table 7.3 Item-person map



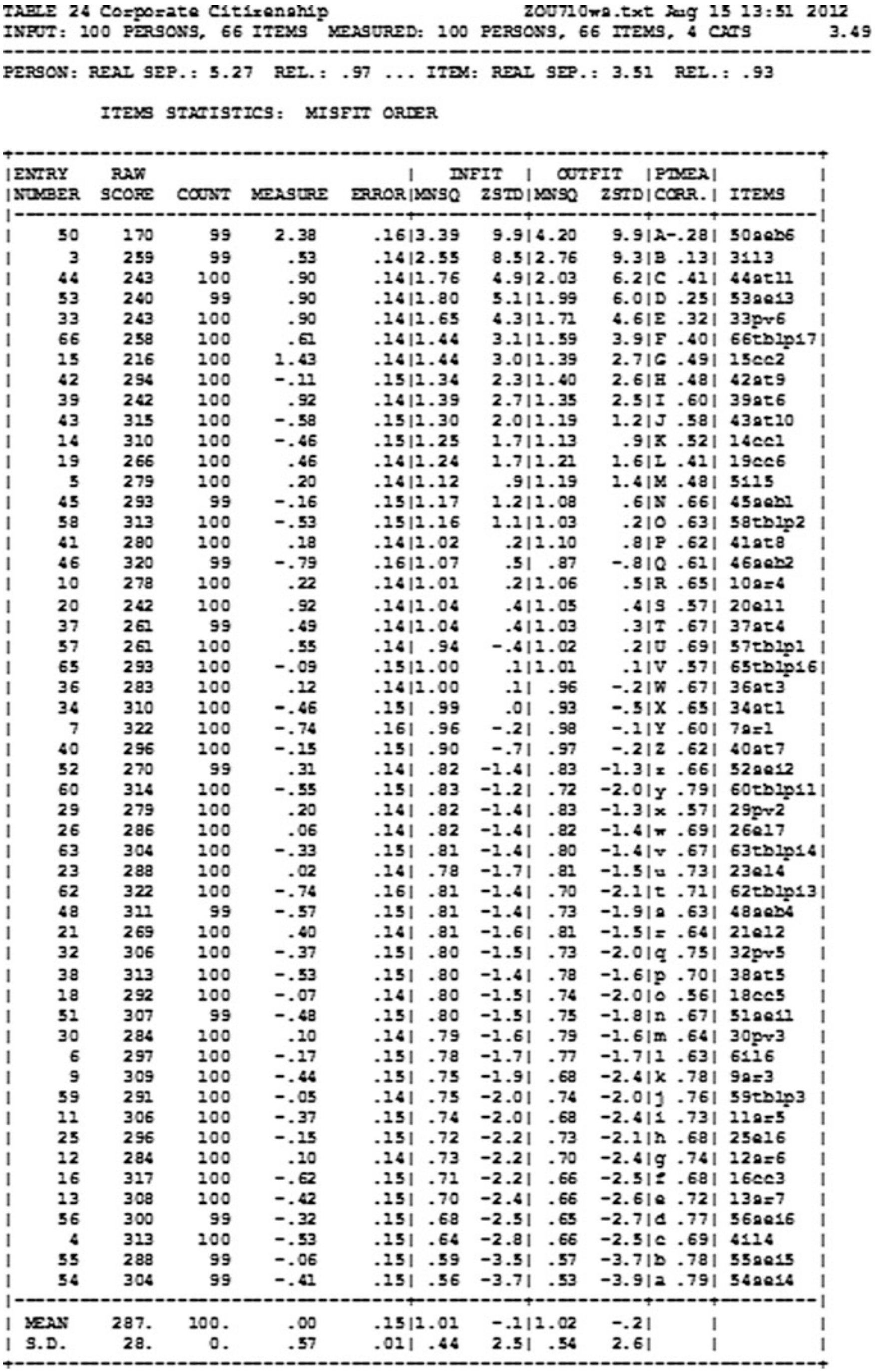
```
TABLE 23 Corporate Citizenship                ZOU710wa.txt Aug 15 13:51 2012

INPUT: 100 PERSONS, 66 ITEMS  MEASURED: 100 PERSONS, 66 ITEMS, 4 CATS      3.49
--------------------------------------------------------------------------------

       PERSONS MAP OF ITEMS
          <more>|<rare>
    4           +
                |
                |
             X  |
                |
                |
            XX  |
    3          T+
             X  |
         XXXXX  |
                |
                |  50aeb6
            XX  |
         XXXXX  |
    2        X  +
          XXXX S|
           XXX  |
             X  |
       XXXXXXX  |  15cc2
            XX  |
            XX  |T
    1        X  +
          XXXX  |  20el1     33pv6     39st6     44st11    53ae13
        XXXXXX M|
      XXXXXXXX  |S 3i13      57tblp1   66tblp17  8ar2
            XX  |  19cc6     21el2     37st4
      XXXXXXXX  |  10ar4     24el5     28pv1     52ae12
      XXXXXXXX  |  12ar6     29pv2     30pv3     36st3     41st8     5i15
    0      XXX +M 18cc5     23el4     26el7     2i12      31pv4     35st2
                   49aeb5    55ae15    59tblp3
             X  |  22el3     25el6     40st7     42st9     45aeb1    47aeb3
                   65tblp16  6i16
          XXXX  |  17cc4     56ae16    63tblp14
            XX S|  11ar5     13ar7     14cc1     27el8     32pv5     34st1
                   51ae11    54ae14    64tblp15  9ar3
       XXXXXXX  |S 16cc3     38st5     43st10    48aeb4    4i14      58tblp2
                   60tblp11
           XXX  |  1i11      61tblp12  62tblp13  7ar1
          XXXX  |  46aeb2
   -1        X  +
             X  |T
                |
                |
               T|
             X  |
                |
   -2           +
          <less>|<frequ>
```

construct on a continuum of increasing intensity. Statistically, infit and outfit MNSQ for items should be greater than 0.5 and less than 1.5 for rating scale application. Similarly, MNSQ statistics for Pearson measures should also be less than 1.8. Bond and Fox (2010) provided an elaboration on the infit statistic,

Table 7.4 Item fit order



```
TABLE 24 Corporate Citizenship                         ZOU710wa.txt Aug 15 13:51 2012
INPUT: 100 PERSONS, 66 ITEMS  MEASURED: 100 PERSONS, 66 ITEMS, 4 CATS          3.49
------------------------------------------------------------------------------------
PERSON: REAL SEP.: 5.27  REL.: .97 ... ITEM: REAL SEP.: 3.51  REL.: .93

        ITEMS STATISTICS:  MISFIT ORDER

+-----------------------------------------------------------------------------+
|ENTRY    RAW                                |   INFIT  |  OUTFIT  |PTMEA|          |
|NUMBER  SCORE  COUNT  MEASURE  ERROR|MNSQ  ZSTD|MNSQ  ZSTD|CORR.| ITEMS    |
|------------------------------------+----------+----------+-----+----------|
|     50    170     99    2.38     .16|3.39   9.9|4.20   9.9|A-.28| 50aeb6   |
|      3    259     99     .53     .14|2.55   8.5|2.76   9.3|B .13| 3i13     |
|     44    243    100     .90     .14|1.76   4.9|2.03   6.2|C .41| 44at11   |
|     53    240     99     .90     .14|1.80   5.1|1.99   6.0|D .25| 53ae13   |
|     33    243    100     .90     .14|1.65   4.3|1.71   4.6|E .32| 33pv6    |
|     66    258    100     .61     .14|1.44   3.1|1.59   3.9|F .40| 66tblpi17|
|     15    216    100    1.43     .14|1.44   3.0|1.39   2.7|G .49| 15cc2    |
|     42    294    100    -.11     .15|1.34   2.3|1.40   2.6|H .48| 42at9    |
|     39    242    100     .92     .14|1.39   2.7|1.35   2.5|I .60| 39at6    |
|     43    315    100    -.58     .15|1.30   2.0|1.19   1.2|J .58| 43at10   |
|     14    310    100    -.46     .15|1.25   1.7|1.13    .9|K .52| 14cc1    |
|     19    266    100     .46     .14|1.24   1.7|1.21   1.6|L .41| 19cc6    |
|      5    279    100     .20     .14|1.12    .9|1.19   1.4|M .48| 5i15     |
|     45    293     99    -.16     .15|1.17   1.2|1.08    .6|N .66| 45aeb1   |
|     58    313    100    -.53     .15|1.16   1.1|1.03    .2|O .63| 58tblp2  |
|     41    280    100     .18     .14|1.02    .2|1.10    .8|P .62| 41at8    |
|     46    320     99    -.79     .16|1.07    .5| .87   -.8|Q .61| 46aeb2   |
|     10    278    100     .22     .14|1.01    .2|1.06    .5|R .65| 10ar4    |
|     20    242    100     .92     .14|1.04    .4|1.05    .4|S .57| 20e11    |
|     37    261     99     .49     .14|1.04    .4|1.03    .3|T .67| 37at4    |
|     57    261    100     .55     .14| .94   -.4|1.02    .2|U .69| 57tblp1  |
|     65    293    100    -.09     .15|1.00    .1|1.01    .1|V .57| 65tblpi16|
|     36    283    100     .12     .14|1.00    .1| .96   -.2|W .67| 36at3    |
|     34    310    100    -.46     .15| .99    .0| .93   -.5|X .65| 34at1    |
|      7    322    100    -.74     .16| .96   -.2| .98   -.1|Y .60| 7ar1     |
|     40    296    100    -.15     .15| .90   -.7| .97   -.2|Z .62| 40at7    |
|     52    270     99     .31     .14| .82  -1.4| .83  -1.3|z .66| 52ae12   |
|     60    314    100    -.55     .15| .83  -1.2| .72  -2.0|y .79| 60tblpi11|
|     29    279    100     .20     .14| .82  -1.4| .83  -1.3|x .57| 29pv2    |
|     26    286    100     .06     .14| .82  -1.4| .82  -1.4|w .69| 26e17    |
|     63    304    100    -.33     .15| .81  -1.4| .80  -1.4|v .67| 63tblpi14|
|     23    288    100     .02     .14| .78  -1.7| .81  -1.5|u .73| 23e14    |
|     62    322    100    -.74     .16| .81  -1.4| .70  -2.1|t .71| 62tblpi13|
|     48    311     99    -.57     .15| .81  -1.4| .73  -1.9|s .63| 48aeb4   |
|     21    269    100     .40     .14| .81  -1.6| .81  -1.5|r .64| 21e12    |
|     32    306    100    -.37     .15| .80  -1.5| .73  -2.0|q .75| 32pv5    |
|     38    313    100    -.53     .15| .80  -1.4| .78  -1.6|p .70| 38at5    |
|     18    292    100    -.07     .14| .80  -1.5| .74  -2.0|o .56| 18cc5    |
|     51    307     99    -.48     .15| .80  -1.5| .75  -1.8|n .67| 51ae11   |
|     30    284    100     .10     .14| .79  -1.6| .79  -1.6|m .64| 30pv3    |
|      6    297    100    -.17     .15| .78  -1.7| .77  -1.7|l .63| 6i16     |
|      9    309    100    -.44     .15| .75  -1.9| .68  -2.4|k .78| 9ar3     |
|     59    291    100    -.05     .14| .75  -2.0| .74  -2.0|j .76| 59tblp3  |
|     11    306    100    -.37     .15| .74  -2.0| .68  -2.4|i .73| 11ar5    |
|     25    296    100    -.15     .15| .72  -2.2| .73  -2.1|h .68| 25e16    |
|     12    284    100     .10     .14| .73  -2.2| .70  -2.4|g .74| 12ar6    |
|     16    317    100    -.62     .15| .71  -2.2| .66  -2.5|f .68| 16cc3    |
|     13    308    100    -.42     .15| .70  -2.4| .66  -2.6|e .72| 13ar7    |
|     56    300     99    -.32     .15| .68  -2.5| .65  -2.7|d .77| 56ae16   |
|      4    313    100    -.53     .15| .64  -2.8| .66  -2.5|c .69| 4i14     |
|     55    288     99    -.06     .15| .59  -3.5| .57  -3.7|b .78| 55ae15   |
|     54    304     99    -.41     .15| .56  -3.7| .53  -3.9|a .79| 54ae14   |
|------------------------------------+----------+----------+-----+----------|
| MEAN    287.   100.      .00     .15|1.01   -.1|1.02   -.2|     |          |
| S.D.     28.     0.      .57     .01| .44   2.5| .54   2.6|     |          |
+-----------------------------------------------------------------------------+
```

"it gives relatively more weight to the performances of person closer to the item value. The argument is that persons whose ability is close to the item's difficulty should give a more sensitive insight into that item's performance."

7.7 Principal Component

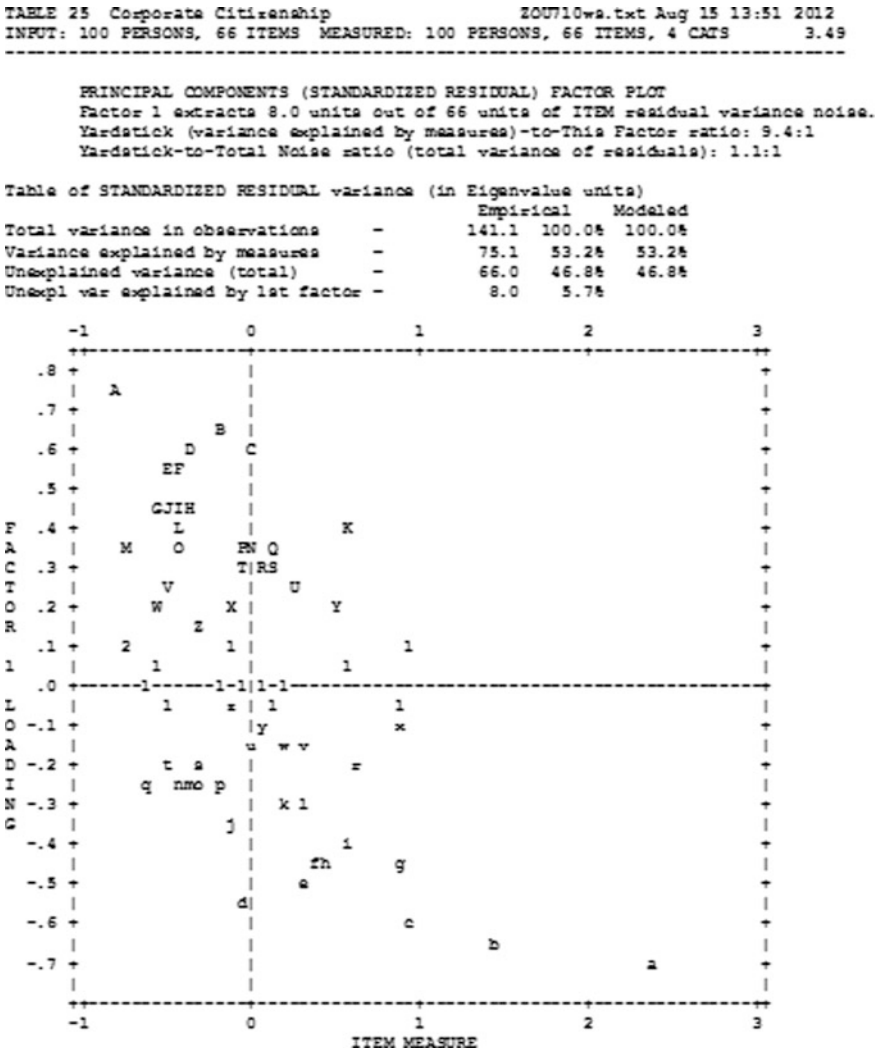
Factor analysis technique is another way to detect important deviations from the unidimensionality of Rasch modelling (Wright 1996). Here, Table 7.5 illustrates the variance explained by measure as 53.2 % which demonstrated that the items were able to endorse the corporate citizenship practise among the senior managers of the public listed companies. The analysis also elucidates the need for the researchers to explore further the unexplained variance results. This next section extends an explication on this.

7.8 Understanding Misfits

Based on the diagnostics applied, (i) summary statistics, (ii) item fit order, (iii) item polarity map, (iv) item-person map and (v) principal component, the main findings demonstrated that the Rasch criteria were met. Further, the fit indices also helped the study to confirm that the assumption of unidimensionality was held up empirically. In addition, the high item reliability index of 0.93 indicated confidence in the replicability of item placement across other samples (Bond and Fox 2010).

At the same time, it is pertinent in Rasch to analyse misfits. In this study, some items (seb6, il3, st11, sei3 and pv6) showed a misfit to the Rasch model. According to Bond and Fox (2010), misfitting items should be investigated, and if there were any theoretical basis for their inclusion, they should be examined further. Consequently, two components that required further probing were 'level of stakeholder engagement influenced by stakeholder pressure' and 'level of stakeholder prioritisation' on Orang Asli (Indigenous Peoples). Upon critical examination of the results, it was evident that the misfit reported by Rasch was attributed to the mixed responses on these items across the polytomous categories from 1 (non-significant) to 4 (very significant). Following a deliberation and literature review on this finding, the mixed responses are reflective of the actual state of affair and perception on these two components among companies in Malaysia (SUHAKAM 2011; United Nations 2009). It was evident that both Indigenous peoples (Orang Asli) and stakeholder pressures were not regarded as serious concerns by many Malaysian companies, although recognition on these components is gaining prominence among some industries such as plantation, real estate and property in Malaysia. In countries such as Australia, the results would be different, as most industries are obliged to address Indigenous peoples concerns in line with statutory commitments as well as high stakeholder pressure from various stakeholder groups. In addressing

Table 7.5 Principal component



```
TABLE 25  Corporate Citizenship                  ZOU710ws.txt Aug 15 13:51 2012
INPUT: 100 PERSONS, 66 ITEMS  MEASURED: 100 PERSONS, 66 ITEMS, 4 CATS       3.49
--------------------------------------------------------------------------------

       PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT
       Factor 1 extracts 8.0 units out of 66 units of ITEM residual variance noise.
       Yardstick (variance explained by measures)-to-This Factor ratio: 9.4:1
       Yardstick-to-Total Noise ratio (total variance of residuals): 1.1:1

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                             Empirical    Modeled
Total variance in observations     -        141.1  100.0%  100.0%
Variance explained by measures     -         75.1   53.2%   53.2%
Unexplained variance (total)       -         66.0   46.8%   46.8%
Unexpl var explained by 1st factor -          8.0    5.7%

      -1               0               1               2                3
      ++---------------+---------------+---------------+---------------++
   .8 +                |                                                +
      |   A            |                                                |
   .7 +                |                                                +
      |             B  |                                                |
   .6 +          D     C                                                +
      |       EF       |                                                |
   .5 +                |                                                +
      |      GJIH      |                                                |
F  .4 +         L      |        K                                       +
A     |    M    O     PN Q                                              |
C  .3 +               T|RS                                              +
T     |       V        |   U                                            |
O  .2 +      W      X  |       Y                                        +
R     |           Z    |                                                |
   .1 +    2         1 |              1                                 +
1     |      1         |        1                                       |
   .0 +------1------1-1|1-1---------------------------------------------+
L     |       1     z  | 1           1                                  |
O -.1 +                |y           x                                   +
A     |                u  w v                                           |
D -.2 +       t  a     |         r                                      +
I     |      q  nmo p  |                                                |
N -.3 +                |  k l                                           +
G     |              j |                                                |
  -.4 +                |        i                                       +
      |                |     fh      g                                  |
  -.5 +                |    e                                           +
      |               d|                                                |
  -.6 +                |              c                                 +
      |                |                      b                         |
  -.7 +                |                                     a          +
      |                |                                                |
      ++---------------+---------------+---------------+---------------++
      -1               0               1               2                3
                                 ITEM MEASURE
```

the limitation, the embedded mixed methods design (Creswell 2009) was adopted. This means one data set provided a supportive or complementary, secondary role in a study that based primarily on another data set (Creswell and Clark 2007). A review of literature revealed that while different terms or classification has been used for this design, the embedded or concurrent nested method has been adopted in a number of research disciplines, namely, social and behavioural research

(Tashakkori and Teddlie 2003), evaluation (Greene and Caracelli 1997), nursing and educational research (Tashakkori and Teddlie 1998; Creswell 2003). Finally, further research that includes bigger sample size and even cross-border studies on these elements would lend useful insights into theory and practise.

7.9 Conclusion

The application of the Rasch measurement model in corporate citizenship research should be given more consideration by researchers. Its relevance and currency in mitigating the limitations and challenges faced in the field of social science, especially in relation to establishing psychometrically sound scales to measure the perception of individuals, groups and cross cultural analysis, as well as, establishing unidimensionality through its rigorous process deserves serious attention and should be explored more extensively in our field of research in Malaysia.

References

Andrich, D. (1988). *Rasch models for measurement. Series: Quantitative applications in the social sciences.* 68. USA: Sage Publications Inc.

Antony, J. P., & Bhattacharyya, S. (2010). Measuring organizational performance and organizational excellence of SMEs—Part 1: A conceptual framework. *Measuring Business Excellence, 14*(2), 3–11.

Bamber, C. J., Sharp, J. M., & Hides, M. T. (2000). Developing management systems towards integrated manufacturing: A case study perspective. *Integrated Manufacturing Systems, 11*(7), 454–461.

Bond, T. G., & Fox, C. M. (2010). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge Taylor and Francies Group.

Castka, P., Bamber, C. J., Bamber, D. J., & Sharp, J. M. (2004). Case Study: Integrating Corporate Social Responsibility (CSR) into ISO Management Systems – in search of a feasible CSR management system framework. *The TQM Magazine, 16*(3), 216–224.

Creswell, J. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks: Sage.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative and mixed methods approaches*. CA: Sage Publications Inc.

Creswell, J. W., & Clark, V. L. P. (2007). *Designing and conducting mixed methods research*. CA: Sage Publications Inc.

Duncan, O. D. (1984). Rasch measurement: Further examples and discussion. In Turner C. F., & Martin E. (Eds.) *Surveying Subjective Phenomena* (Vol. 2, Chapter 12, pp. 367–403). New York: Russell Sage Foundation.

Fischer, A. R. H., Frewer, L. J., & Nauta, M. A. (2006). Toward improving food safety in the domestic environment: A multi-item Rasch scale for the measurement of the safety efficacy of domestic food-handling practices. *Risk Analysis, 26*, 1323–1328.

Frankental, P. (2001). Corporate Social Responsibility – a PR invention? *Corporate Communications: An International Journal, 6*(1), 18–23.

Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. *New Directions for Evaluation*, 5–17. doi: 10.1002/ev.1068.

Hermel, P., & Ramis-Pujol, F. (2003). An evolution of Excellence some main trends. *The TQM Magazine, 15*(4), 23–40.

Islam, A. Y. M. (2011). *Online database adoption and satisfaction model – factors influencing the adoption of an online database among students and their satisfaction in using it*. LAP Lambert Academic Publishing.

Jung, J. Y., & Hong, S. (2008). *Organizational citizenship behaviour (OCB), TQM and performance at the maquiladora, Emerald 25*. Emerald Group Publishing Limited.

Mcintosh, M., Thomas, R., Leipziger, D., & Coleman G. (2003). *Living Corporate Citizenship: Strategic Routes to Socially Responsible Business*. London, UK: Prentice Hall, Financial Times (Pearson Education Limited).

Mcintosh, K., Earleywine, M., & Dunn, M. E. (2006). Alcohol expectancies for social facilitation. *Journal of the American Medical Association, 270*, 2207–2212.

Olsen, J. E. (2002). Global Ethics and the Alian Tort Claims Act: A Summary of Three Cases with the Oil and Gas Industry. *Management Decision (Emerald publishing) ©MCB UP Ltd, 40*(7), 720–724.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the hospital anxiety and depression scale (HADS). *The British Journal of Clinical Psychology., 46*, 1–18.

Pillai, K. V. K. (2013). The influence of corporate conduct and stakeholder empowerment on corporate citizenship in Malayisa: A mixed methods research. *Doctoral dissertation*. Open University Malaysia.

Searing, L. M. (2008). *Family functioning scale validation: A Rasch analysis. Dissertation*. Chicago: University of Illinois.

Smith, R. (1993). Guessing and the Rasch model. *Rasch Measurement Transactions, 6*(4), 262–263.

Smith, R. M. (1999). *Rasch measurement models: Interpreting WINSTEPS/Bigsteps and Facets output*. Morgan Hill, CA: JAM Press.

SUHAKAM. (2011). *Suhakam annual report 2011*. Kuala Lumpur: Human Rights Commission of Malaysia.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. London: Sage Publications.

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social behavioral research*. Thousand Oaks: Sage

United Nations. (2009). *State of the world's indigenous peoples*. Department of Economic and Social Affairs. New York: United Nations Publications.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281–288.

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

Wright, B. D., & Masters, G. N. (1982). *Rating Scale analysis*. Chicago, IL: MESA Press.

Chapter 8
Fluency: A Critically Important Yet Undervalued Dimension of L2 Vocabulary Knowledge

Junyu Cheng, Joshua Matthews and John Mitchell O'Toole

Abstract Measurements of vocabulary knowledge are powerful indicators of language proficiency, as they correlate with other language skills (Milton 2013; Stæhr 2009). It is evident to all L2 educators that L2 vocabulary knowledge is an essential component of L2 competency. The underlying constructs of vocabulary knowledge are generally accepted as consisting of multiple dimensions (Milton 2013). Among these knowledge dimensions, vocabulary fluency, or the ability to access and apply vocabulary knowledge under time constraints, is arguably one of the most important. This is especially the case in regard to the ability to apply existing L2 vocabulary knowledge in a communicatively competent manner. This paper asserts that the ability to apply vocabulary knowledge in a fluent manner has been systemically undervalued in contemporary L2 educational contexts. In support of this assertion, empirical data drawn from the results of two equivalent vocabulary tests administered among a cohort of 113 tertiary-level L2 learners are presented. Results showed significantly lower mean vocabulary scores on the fluency dependent (FD) test than those attained by the cohort on the fluency non-dependent (FND) test. The implications of this significant discrepancy between these two measures of vocabulary knowledge are discussed. The paper concludes that the cohorts' FD knowledge of high frequency and academic words was significantly lower than their FND knowledge of those same categories of words. These results encourage an invigoration of teaching, testing, and

J. Cheng (✉)
School of Foreign Languages, Southeast University, 2 Si Pai Lou,
Nanjing 210096, P.R. China
e-mail: chjy@seu.edu.cn

J. Matthews · J.M. O'Toole
School of Education, University of Newcastle, Callaghan, Australia
e-mail: joshua.matthews@uon.edu.au

J.M. O'Toole
e-mail: mitch.otoole@newcastle.edu.au

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_8

learning approaches that emphasize the development of FD word knowledge. Practical suggestions on improving the vocabulary fluency of English as a second language students are provided.

Keywords Academic vocabulary · High frequency vocabulary · L2 vocabulary knowledge · Vocabulary fluency · Vocabulary testing

8.1 Introduction

8.1.1 The Importance of Vocabulary Knowledge

Vocabulary knowledge is one of the most crucial factors in the development of second language (L2) acquisition and language performance (Hulstijn 2002; Milton 2013). Vocabulary knowledge holds a special status in relation to language learning and proficiency in that it is at the level of the word that stable formal and semantic associations occur (Hulstijn 2002). The importance of a pedagogical focus at the level of the word is also highlighted by the fact that vocabulary knowledge is a powerful indicator of language proficiency, being well correlated with the macroskills of reading, writing, listening, and speaking (Milton 2013; Nation 2001; Qian 2002; Stæhr 2008, 2009). For these reasons, the measurement and analysis of vocabulary knowledge provide a valuable access point from which to investigate the L2 proficiency of cohorts of L2 learners.

Strategic measurement of different aspects of vocabulary knowledge provides language teachers with a robust means by which to diagnose specific areas of weakness in the language proficiency among cohorts of L2 learners. In the immediate term, the identification of any systemic weaknesses in vocabulary knowledge can be useful in drawing the teacher's attention to potential deficits in day-to-day teaching emphases. More broadly, identifying areas of weakness in vocabulary knowledge can be used as an empirical basis from which to guide the design and implementation of pedagogical interventions aimed at remedying systemic weaknesses in teaching and learning programs.

Prior to any discussion of the use of vocabulary knowledge and its measurement as a means to guide improved teaching and learning approaches, a careful consideration of the nature of vocabulary knowledge is warranted. Vocabulary knowledge is broadly accepted as being a multidimensional construct consisting of a number of interlocking yet measurably different knowledge domains. Although there is currently no universally accepted framework for vocabulary knowledge, several authors have proposed frameworks that have a number of concordant features (Daller et al. 2007; Henriksen 1999; Meara 1996). The underlying constructs of vocabulary knowledge are generally accepted as consisting of three dimensions that can be broadly defined as vocabulary breadth, vocabulary depth, and vocabulary fluency (Milton 2013).

Vocabulary breadth refers to “the number of words the meaning of which one has at least some superficial knowledge” (Qian 2002, p. 515). Vocabulary breadth “is minimally seen as a basic form-meaning mapping” (Gyllstad 2013, p. 14). Vocabulary breadth deals with the sheer number of words known; however, this dimension of vocabulary knowledge does not provide a means by which to describe the quality of that knowledge. *Vocabulary depth* provides a measure of how well one knows a lexical item (Qian 2002). Vocabulary depth is seen as the aspects of knowing a word which extend beyond the basic capacity of the language learner to establish form-meaning associations (Gyllstad 2013). These may include but are not limited to various aspects of receptive and productive knowledge of word form, meaning, and use (Nation 2001; Qian 2002). *Vocabulary fluency* relates to the ability to access and apply existing word knowledge under time constraints. Daller et al. (2007, p. 8) describe vocabulary fluency as “how readily and automatically a learner is able to use the words they know and the information they have on the use of these words.”

Based on an assumption that vocabulary knowledge entails vocabulary breadth, depth, and fluency, it stands to reason that effective approaches to vocabulary testing and teaching should cater to each of these broad domains of vocabulary knowledge. Surveys of recent literature related to the development of vocabulary knowledge indicate that not all of these types of vocabulary knowledge have been equally emphasized. A relatively strong research emphasis has been allocated to vocabulary breadth and vocabulary depth (Nation 2001, 2006; Qian 2002; Stæhr 2008, 2009). In contrast, the construct of vocabulary fluency has only recently started to receive concerted research effort (Zhang and Lu 2013). It is our experience that vocabulary fluency is also underemphasized in contemporary academic contexts. Specifically, there is a deficit in both pedagogical approaches aimed at improving learners’ ability to fluently use specific vocabulary items and the implementation of testing instruments that measure vocabulary fluency. A central thesis of this paper is that this lack of teaching and testing emphasis on vocabulary fluency has a real and measurable impact on learners’ ability to apply their vocabulary knowledge within time-constrained contexts.

In this paper, empirical data are presented which demonstrates an example of the degree to which measurable differences between a form of fluency dependent (FD) and fluency nondependent (FND) vocabulary knowledge are evident among a cohort of 113 English as a second language (ESL) learners within a Chinese tertiary educational context.

8.1.2 The Importance of Vocabulary Fluency in L2 Development

Although the development of a large vocabulary size is an essential component of L2 vocabulary development, it is seen as merely a solid foundation. To this foundation of knowledge must be added the ability to access and apply vocabulary knowledge in an automatized manner (Hulstijn 2007). An important element of fluent language use is automatic word recognition (in listening and reading) and automatic word retrieval (in speaking and writing).

The importance of vocabulary fluency, or the speed at which existing vocabulary knowledge can be accessed and applied, in large part relates to the finite attentional capacity humans have for processing information. The ability to process linguistic data in an automatic manner affords great benefit in relation to managing the potential cognitive burden imposed by the need to process large amounts of linguistic information in a short period of time. The greater the degree of automaticy in language processing at lower linguistic levels, the more attentional resources language users are able to put toward interpreting meaning at higher or more abstract levels. It is especially the recognition of words that must be automatized so as to free up attentional capacity for the processing of the meaning of the text that is being produced or perceived (Segalowitz and Hulstijn 2005). It is also pointed out that L2 curricula often allow too little time to be devoted to the training of fluency skills (Hulstijn 2007). Fluency-promoting activities, therefore, must have a prominent place in the L2 curriculum with suggestions of 25 % of learning time being allocated to this task (Nation 2001).

Both in the context of L2 performance and in the context of first language (L1), speed, accuracy, and smoothness are the hallmarks of good performance in speech and reading (Berninger et al. 2001; Levelt 1989; Perfetti 1985, 2007).

Taking oral English as the first example, Hulstijn (2007) pointed out that speaking is primarily lexically driven. In general, the lexicon comes first and grammar only second. This is largely true as well for listening, reading, and writing. Therefore, automatic word retrieval initiated by nonverbal thoughts of a speaker have a pivotal function in one's oral English performance.

Segalowitz and Freed (2004, p. 117) hold the view that

> speed and efficiency of L2-specific lexical access and attention control would be related to oral fluency in various ways. For example, being able to access meanings quickly and efficiently should enhance speech rate and reduce hesitations and interruptions that characterize less fluent speech. It was also hypothesized that these cognitive variables might serve as readiness factors for oral gains because to make use of language input the learner has to be able to process that input well. If processing abilities are below some threshold level of readiness, then gains in oral performance may not occur.

Listening, like speaking, is largely a matter of automatic, parallel processing. The lower order processes of word recognition play a crucial role in these automatic processes, as it is at the level of words (i.e., lexemes) that forms are matched with meanings (Hulstijn 2007).

Understanding the meaning of utterances involves many stages (Rost 2002). The most crucial one is the word-by-word understanding of what is being said. Only after one or more words have been identified can the higher order processes begin to operate (Hulstijn 2007).

Hulstijn (2003) points out that L2 learners must get the opportunity to make themselves familiar with the phonetic and phonological properties of the L2, learn large amounts of words, and automatize their ability to recognize words in speech in the training of listening skills. “We may conclude that the acquisition of word recognition skills in listening requires special attention in the L2 curriculum (Hulstijn 2003, p. 420).

Similarly, word recognition is the most important factor in fluent reading (Perfetti 1994).

> When we read, processes of word recognition normally take place automatically. In contrast, children, in initial stages of learning to read, pay so much attention to reading individual words that they have no attention capacity left for the meaning of what they read; when reading the fourth word of a sentence they may have already forgotten the first word (Hulstijn 2003, p. 419).

Fluent reading and listening are characterized by automatic processing at the lower levels of word recognition and sentence parsing, leaving attention capacity free to concentrate on the higher levels of information, that is, on semantics and content (Harrington 2001; Rost 2002).

Writing, with complex problem-solving nature, requires perhaps more attention to the highest levels of information than the other language skills (Hulstijn 2007).

This requirement normally exceeds the attentional capacity of the writer. It is mandatory that word retrieval and spelling consumes relatively little time so as to devote more time to higher order information processing. That is why writers with high verbal ability have been shown to spend more time on text coherence than low verbal-ability writers (Glynn et al. 1982).

To sum up, automatic access and application of vocabulary knowledge have an indispensable function in communicative competent language use. It is noted that language learners often possess word knowledge which is inaccessible under time constrained and thus fluency dependent contexts (Goh 2000).

8.1.3 Vocabulary Fluency as an Undervalued Construct in the Chinese Tertiary L2 Educational Context

Despite the acknowledgment that vocabulary is a critically important component of effective language education, such insights are yet to make a significant impact on language teaching (Milton 2013; Schmitt 2008). Indeed, over the past half century or more, vocabulary as an individual subject has been relegated to an adjunct to subjects which emphasize the four macroskills of reading, writing, listening, and speaking (Milton 2013).

Within the Chinese educational context, research regarding vocabulary size has been a strong feature since the 1980s. A majority of the studies have focused on correlating vocabulary size and language competencies. Many Chinese researchers, such as Gui (1985), Yu (1991), Zhou (2000), Deng (2001), Shao (2002) and Lu (2004), have invested considerable research effort in studies of relationship between vocabulary size and language competencies. The dimension of vocabulary depth has also been investigated, but such research to date is relatively scarce. When it comes to how readily and automatically a learner is able to use the words they know, little attention has been paid to this issue. To date research has largely focused on the use of words in writing, with little research effort directed towards word use during listening, reading and speaking.

In the Chinese tertiary L2 educational context, the requirement for vocabulary breadth of College English Curriculum Requirements is 4,795 words for basic level, 6,395 words for intermediate level, and 7,675 words for advanced level. In the view of Cai (2012), this requirement can neither connect with secondary school English teaching reasonably nor reflect the actual situation of vocabulary breadth of college students, thus causing ineffective teaching, failure to achieve the objectives of the College English Curriculum Requirements, and hindering the improvement in English teaching. Cai suggested that the requirement of vocabulary size should be increased. Cai (2001) also asserted that there are many repetitions in the textbooks of tertiary, secondary, and primary education, and the requirement of vocabulary size is small, causing complacency and sluggishness in English learning.

As for vocabulary depth, Liu (2002) found that students acquire receptive knowledge much better than productive knowledge, and negative mother language transfer influences vocabulary depth. Zhang (2006) claimed that vocabulary teaching in class is limited in scope, and of poor quality and depth.

As for teaching methods, teachers teach vocabulary based on their personal experience without a systematic pedagogy. Zhang (2005) found that many teachers ask students to learn vocabulary by themselves and supervise their self-study through quizzes. Pu (2003) asserted that in spite of knowing the significance of vocabulary in English learning, teachers cannot find an effective method and approach to teach vocabulary.

From the perspective of vocabulary learning, learners learn vocabulary only with the purpose of passing exams. Therefore, they mainly enlarge their vocabulary through mechanical recitation and acquire only Chinese meaning of the vocabulary. They do not know how to use the words, let alone internalizing the vocabulary knowledge. Some students cannot understand what the passage conveys although they know every word. Some students find it is really hard to remember a word, and they cannot use it although they did know a word. Zhu (1994) claimed that the lack of vocabulary fluency test leads to the slow improvement in language ability after grasping a certain amount of vocabulary.

In actual L2 teaching contexts, teaching of vocabulary as an individual subject is not encouraged in many tertiary institutions. Vocabulary instruction is often labeled as outdated, traditional, and not communicative in orientation. In language proficiency tests, the provision of a separate section for testing vocabulary knowledge has not been a prominent feature for some time. Although vocabulary is still a component of some low-stake tests, it is dying out gradually from present classroom teaching and testing regimes.

In the limited testing practice of vocabulary, vocabulary tests measure only the breadth and depth of vocabulary knowledge. As the construct of vocabulary fluency is hardly involved in present English testing, this construct is seriously undervalued and far less strongly encouraged than other constructs of language competence. Compounding the difficulty in these issues is the difficulty associated with developing vocabulary fluency within formal instructional contexts. These difficulties have resulted in vocabulary fluency being completely absent from our English curriculum and largely ignored in existing L2 education context.

8.2 Research Questions

To test the hypothesized proficiency deficit in the ability of students' to apply their vocabulary knowledge under time constraints, the following research questions will be addressed:

8.2.1 Research Question 1

Is there a significant difference between the FD and FND vocabulary knowledge among L2 learners in our research context?

Our hypothesis in relation to this research question stems from our belief that within our research context, the L2 vocabulary fluency is less strongly emphasized in testing and teaching than is FND vocabulary knowledge. We therefore hypothesize that FD vocabulary tests scores will be significantly less than FND vocabulary test scores.

8.2.2 Research Question 2

Is there a significant difference between the FD and FND vocabulary knowledge for high frequency words among L2 learners in our research context?

This research question will investigate the FD and FND vocabulary knowledge of high-frequency vocabulary items. These words are important as a high proportion of written and spoken texts are comprised of a relatively small but very frequently used group of words (Nation 2001).

8.2.3 Research Question 3

Is there a significant difference between the FD and FND vocabulary knowledge for academic words among L2 learners in our research context?

This research question will investigate the FD and FND vocabulary knowledge of academic words (Coxhead 2000). These words are those that are less frequent than the high-frequency words but still comprise a significant proportion of the language used in academic discourse (Nation 2001). This category of words is likely to be of great significance to tertiary-level students seeking to use English language skills in specific professional and academic contexts in the future.

Our hypotheses in relation to research questions 2 and 3 are in line with our first hypothesis. Our assertion is that fluency is a dimension of vocabulary knowledge that has been systemically overlooked in vocabulary learning, and therefore we expect that FD knowledge for both high-frequency and academic words to be significantly less than that for FND vocabulary knowledge for words in those same categories.

8.3 Methodology

8.3.1 Participants

A total of 113 first-year undergraduate students made up the participants involved in this study. These participants were members of three separate classes enrolled in the same English language course studying at a large Chinese university. The combined group comprised 65 males (58 %) and 48 females (42 %). The mean self-reported duration of English language study was 9.4 years (*SD* = 2.4, minimum 4 years and maximum 18 years). The mean age of the participants was 18.5 years. All were highly proficient speakers of Mandarin Chinese, with over 92 % reporting Mandarin Chinese as their mother tongue.

8.3.2 Instruments

Two aspects of vocabulary knowledge, FND knowledge, and FD knowledge were measured for all 113 members of the research project. Each of these aspects of vocabulary knowledge was measured using a different testing instrument, each of which will be described subsequently. The target words for each of the tests were drawn from two important categories of target words: high-frequency words and academic words. High-frequency words are those which comprise a large proportion of spoken and written discourse, and academic words are those words which are typical of academic discourse (Coxhead 2000).

FND vocabulary knowledge was measured using the high-frequency and academic word-level sections of a previously validated version of the Vocabulary Levels Tests. A total of 22 items were used with each item consisting of three target words. The test consisted of 10 items consisting of 30 high-frequency target words and 12 items consisting of 36 academic target words. As can be seen from the example item shown in Fig. 8.1, the test format involves matching the target words

words	**meanings**
1 copy	
2 event	_____ end or highest point
3 motor	_____ this moves a car
4 pity	_____ thing made to be like another
5 profit	
6 tip	

Fig. 8.1 Example item for fluency non-dependent vocabulary knowledge

with their correct meaning. Reflecting on our working definition of fluency, the degree to which lexical knowledge can be accessed and applied under time constraints, we can see that this test format isn't strongly dependent on vocabulary fluency.

The test taker has the opportunity to consider both the form of the target words and the meaning of the words over a relatively unconstrained duration of time. The test taker can consider and reconsider the content of the item, thus engaging explicit knowledge about those target words in an effort to select the correct answer. Although the Vocabulary Levels Tests format has proven itself to be a very useful measure of vocabulary size in previous studies, it certainly doesn't strongly tap into the definition of vocabulary fluency that has been outlined earlier.

A construct of *FD vocabulary knowledge* was operationalized through the use of a partial dictation test. The format of partial dictation involves the presentation of a written contextual sentence with the target word missing. The test is accompanied by aural stimulus that contains the entire sentence including the target word. Test takers must listen and transcribe in writing the target word as it is heard. This partial dictation test contained 60 items. The first 32 target words were high-frequency words, and the remaining items contained 28 academic words. An example of the written stimulus seen by the test taker is:

This country has a good ……………..………… system.

Test takers listen to the stimulus sound file and must transcribe the word in written form. The aural stimulus for each item is heard only once, and there is a set time period of four seconds between the end of one sentence and the start of another. Thus, the test taker is required to attend to the stimulus material and recognize the target word and apply this lexical knowledge in a time-constrained manner.

Care was also taken to ensure that the written contextual sentence of the partial dictation items could not be used to systematically guess the target word without the assistance of the aural stimulus. This was achieved by piloting the written components of the tests with native speakers. It was determined that without the assistance of the aural stimulus, the target words for the partial dictation tests could not be reliably and systematically guessed from the context.

8.3.3 Procedures

The tests were administered on the same day in each of the participants' normal language classrooms under strict test conditions. All students were given the opportunity to complete the FND test in a time frame of approximately 20 min. All students were able to complete the test within that time frame. Aural stimulus for the FD test was administered via high-quality audio speakers. These tests were also completed within a time frame of approximately 20 min.

8.3.4 Data and Analysis

FND tests (Vocabulary levels test) were marked using an objective marking scheme. The subjectivity associated with the scoring of the FD test (partial dictation test) was overcome using a structured marking rubric. This rubric had been piloted in a previous study (Matthews and O'Toole 2013) and was determined to be an effective marking procedure to ensure minor spelling errors present in the answers did not affect the validity and reliability of the marking approach.

For each of the 113 participants involved, the total marks for both the FND and the FD tests were established and expressed as a percentage of the maximum possible scores. Additionally, for each test, the scores attained for each of the high-frequency words and academic words were calculated and expressed as a percentage.

Prior to conducting the analysis, the assumption of normally distributed data was interrogated by calculating the skew and kurtosis levels present in the data. A range of these values was estimated at between −1.21 and 1.40 which resides within acceptable levels (Posten 1984). The correlation between the FD and the FND total test scores was estimated as $r = 0.70$, $p < 0.01$, suggesting that the use of paired sample *t*-test was an appropriate analytical tool in this instance.

8.4 Results

8.4.1 Overview

See Fig. 8.2 and Table 8.1.

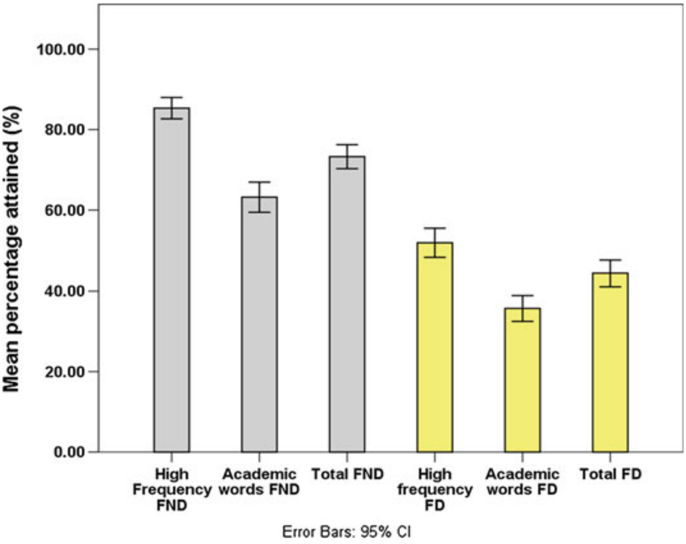


Fig. 8.2 A comparison of mean scores attained for FD and FND test scores

Table 8.1 Descriptive statistics for FD and FND test scores

Test type	*N*	Minimum (%)	Maximum (%)	Mean (%)	*SD*
High frequency words—fluency non-dependent (FND)	113	16.67	100.00	85.33	14.20
Academic words—fluency non-dependent (FND)	113	11.11	94.44	63.20	20.07
Total words—fluency non-dependent (FND)	113	16.67	96.97	73.26	16.00
High frequency words—fluency dependent (FD)	113	1.56	87.50	51.88	19.44
Academic words—fluency dependent (FD)	113	0.00	73.21	35.69	17.15
Total words—fluency dependent (FD)	113	2.50	78.33	44.32	17.46

8.4.2 Research Question 1

Is there a significant difference between the FD and FND vocabulary knowledge among L2 learners in our research context?

To test the hypothesis that FD total test score ($M = 44.32$ %, $SD = 17.46$) was significantly less than FND total test scores ($M = 73.26$ %, $SD = 16.00$), a paired-samples *t*-test was conducted to compare the mean scores attained.

The null hypothesis that there was no significant difference between the FD and the FND total test scores was rejected, t (112) = 23.60, $p < 0.01$. Thus, our first hypothesis was confirmed in that the FD vocabulary knowledge was significantly lower than that of the FND vocabulary knowledge.

8.4.3 Research Question 2

Is there a significant difference between the FD and FND vocabulary knowledge for high-frequency words among L2 learners in our research context?

To test the hypothesis that FD high-frequency test scores (*M* = 51.88 %, *SD* = 19.45) were significantly less than FND high-frequency test scores (*M* = 85.34 %, *SD* = 14.20), a paired-samples *t*-test was conducted. The null hypothesis that there was no significant difference between the FD and the FND high-frequency word test scores was rejected, t (112) = 22.20, $p < 0.01$. Thus, our second hypothesis was also confirmed as it was found that the FD high-frequency vocabulary knowledge was significantly lower than that of the FND high-frequency vocabulary knowledge.

8.4.4 Research Question 3

Is there a significant difference between the fluency dependent and fluency non-dependent vocabulary knowledge for academic words among L2 learners in our research context?

To test the hypothesis that FD academic word test scores (*M* = 35.70 %, *SD* = 17.15) were significantly less than FND academic word test scores (*M* = 63.20 %, *SD* = 20.08), a paired-samples *t*-test was again conducted. The null hypothesis that there was no significant difference between the mean FD and the mean FND academic word test scores was rejected, t (112) = 17.51, $p < 0.01$. Thus, our third hypothesis was confirmed, as it was found that the FD academic vocabulary knowledge was significantly lower than that of the FND academic vocabulary knowledge.

8.5 Discussions and Implications

We are in agreement with Nation (Nation 2001) who for some time has emphasized fluency as a fundamentally important strand of effective L2 learning. Our findings provide empirical evidence which highlights the significant discrepancy between the vocabulary knowledge that learners possess and that which they are able to apply productively under time constraints. This finding has a number of implications of relevance to the L2 learners in our research context and to those within similar language learning contexts.

8.5.1 Implications in Relation to the Communicative Competence of Language Learners

First, a deficit in the ability to apply vocabulary knowledge in a fluent manner, despite having knowledge of that vocabulary in a fluency non-dependent capacity, has significant implications for the communicative competence of learners. A central goal of any language learning program should include encouraging the ability to use that language knowledge in real-world communicative contexts. As has been outlined earlier in this paper, the skills of speaking and listening are particularly dependent on the ability to access and apply vocabulary knowledge under time-constrained conditions. It would follow then that the construct of vocabulary fluency plays an important role in the ability to interact in an interpersonal manner using spoken language. If learners have a deficit in the ability to apply the vocabulary knowledge that they have in the skills of speaking and listening, then it stands to reason that this type of L2 linguistic deficit represents a pressing concern for language educators.

8.5.2 Implications in Relation to the Predictive Power of Vocabulary Tests and the Effect of Wash-Back on Teaching and Learning

Vocabulary tests that do not tap into constructs of vocabulary fluency are likely to be of limited value in estimating test takers' ability to use known vocabulary items in communicative contexts. Although vocabulary tests that do not measure the construct of vocabulary fluency are likely to have validity in their ability to determine which words a learner knows, it is likely that such test will have questionable validity in relation measuring which of those known words a learner is able to access and apply in a fluent and time-constrained manner.

It is our view that without a systematic change to the way vocabulary is tested, it is likely that the magnitude of the discrepancy between FD and FND vocabulary knowledge evident in the results of this study will persist in future cohorts of language learners in our research context. Implementing vocabulary tests that measure not only the breadth and depth of vocabulary knowledge but also the fluency with which a given word can be accessed and produced are central to our recommendations. Implementing well-researched and developed testing approaches which tap into the construct of vocabulary fluency are likely to have positive wash-back effects of teaching and learning behavior. Implementing tests that tap vocabulary fluency would be a functionally effective way to mandate an increased focus on not only the number of words which are known but also the speed at which those words can be applied.

Although this study applied partial dictation as a means by which to operationalize vocabulary fluency, it is acknowledged that more sophisticated methods of testing vocabulary fluency are possible in the future. Of particular importance in this regard is to apply the potential held by computer-assisted testing in the development of more effective vocabulary fluency tests. The key advantages of computer-assisted testing is that the computer-mediated environment enables the establishment and measurement of variables related to fluent vocabulary access and application, the most important of these being time. Of course, computer-assisted testing also offers the convenience of large-scale and relatively autonomous test administration and scoring. It is suggested here that the concurrent development of computer-assisted learning and testing approaches aimed at developing and measuring vocabulary fluency represent a very valuable area for future research. Ideally, implementing such learning and testing regimes would have a measurable impact on reducing the FD and FND vocabulary knowledge discrepancies observed within cohorts of language learners such as those involved in the present study. Long-term implementation of computer-assisted language learning approaches which have been shown to improving forms of FD vocabulary are of strong interest in this regard (Matthews et al. 2014).

8.5.3 Implications in Relation to the Limited FD Vocabulary Knowledge of High Frequency and Academic Words

The difference between the FD and the FND vocabulary knowledge is evident within both the high-frequency and academic word categories. This finding has a number of significant implications. First, if the results of this study are an accurate predictor of other similar learning contexts, it would seem that although high-frequency vocabulary may be relatively well known in the FND domain, the same vocabulary knowledge would appear to be relatively inaccessible under time-constrained conditions. With knowledge of high-frequency vocabulary holding such an important position in L2 proficiency development, it would seem that a concerted effort to devise approaches to improve the vocabulary fluency of these high-frequency words is of very high importance. A consideration of the finding that conversation typically consists of approximately 90 % high-frequency words emphasizes the need to consider the implications of the problem of relatively poor FD high-frequency vocabulary knowledge (Nation 2001). Without a high level of vocabulary fluency for the most frequent words in the target language, it is clear that institutional-wide goals to improve the communicative competence of L2 learners are likely to be difficult to attain.

Improving the degree to which academic words can be fluently used is also likely to have an impact on the degree to which learners can adroitly interact communicatively in academic discourses. This goal may not be as immediately important as the ability to fluently apply high-frequency vocabulary. However, with the assumption that most tertiary level-language learners are likely to benefit if they possess the ability to fluently use an L2 to negotiate academic and professional contexts, it is a goal nonetheless of significant value.

References

Berninger, V., Abbott, R., Billingsley, F., & Nagy, W. (2001). Processes underlying timing and fluency of reading: Efficiency, automaticity, coordination, and morphological awareness. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 383–414). Baltimore: York Press.

Cai, J. (2001). A new approach to the teaching of College English. *Foreign Language World, 85*, 73–77.

Cai, J. (2012). The constraints on and necessity for upgrading college English vocabulary requirement. *Journal Of PLA University of Foreign Languages, 35*(1), 48–53.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 1–32). Cambridge: Cambridge University Press.

Deng, S. (2001). On the measurement of english vocabulary. *Foreign Language Teaching and Research, 33*(1), 57–62.

Duan, S. (2009). L2 lexical competence and its assessment. *Journal Of PLA University of Foreign Languages*, *32*(2), 51–54.

Glynn, S., Britton, B., Muth, D., & Dogan, N. (1982). Writing and revising persuasive documents: Cognitive demands. *Journal of Educational Psychology, 74*(4), 557.

Goh, C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*(1), 55–75.

Gui, S. (1985). Investigation and analysis of English vocabulary of English majors in China. *Modern Foreign Languages, 1985*(1), 1–6.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective—challenges and potential solutions. In *EUROSLA monographs series 2 L2 vocabulary acquisition, knowledge and use* (pp. 11–28).

Harrington, M. (2001). Sentence processing. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 91–124). New York: Cambridge University Press.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*, 303–317.

Hulstijn, J. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research, 18*(3), 193–223.

Hulstijn, J. (2003). Connectionist Models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning, 16*(5), 413–425.

Hulstijn, J. (2007). Psycholinguitisc perspectives on second language acquisition. In J. Cummins & C. Davidson (Eds.), *The international handbook on English language teaching* (pp. 701–713). Norwell: Springer.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge: Bradford.

Liu, S. (2002). Dimension development and acquisition patterns of L2 word knowledge. *Journal Of PLA University of Foreign Languages, 25*(2), 66–69.

Lu, C. (2004). Vocabulary size and its influence on English achievement as well as its relationship to depth of lexical knowledge. *Foreign Language Teaching and Research, 36*(2), 116–123.

Matthews, J., Cheng, J., & O'Toole, J. M. (2014). Computer-mediated input, output and feedback in the development of L2 word recognition from speech. *ReCALL*, 1–19. doi:10.1017/S0958344014000421.

Matthews, J., & O'Toole, J. M. (2013). Investigating an innovative computer application to improve L2 word recognition from speech. *Computer Assisted Language Learning*, 1–19. doi:10.1080/09588221.2013.864315.

Meara, P. (1996). The dimension of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Milton, J. (2013). L2 vocabulary acquisition, knowledge and use. *EUROSLA Monographs Series, 2*, 57–78.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59–82.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383.

Perfetti, C. A. (1985). Reading skills. *Psychiatry, 50*, 1125–1129.

Perfetti, C. (1994). Psycholinguistics and reading ability. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 849–894). San Diego, CA: Academic Press.

Posten, H. (1984). Robustness of the two-sample t-test. In D. Rasch (Ed.), *Robustness of statistical methods and nonparametric statistics* (pp. 92–99). Berlin: Springer.

Pu, J. (2003). Colligation, collocation and chunk in vocabulary teaching and learning. *Foreign Language Teaching and Research, 35*(6), 438–445.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536.

Rost, M. (2002). *Teaching and researching listening.* Harlow: Longman.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363.

Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition, 26*(02), 173–199.

Segalowitz, N., & Hulstijn, J. (2005). Automaticity in second language learning. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press.

Shao, H. (2002). Study on English vocabulary proficiency of Chinese Normal College students. *Foreign Language World, 90*(4), 61–66.

Stæhr, L. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal, 32*(2), 139–152.

Stæhr, L. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*(4), 577–607.

Yu, A. (1991). Vocabulary training with different goals for trainees. *Foreign Language Teaching and Research, 01*, 42–47.

Zhang, P. (2006). Ten year overview of second language vocabulary acquisition research. *Foreign Languages and Their Teaching, 207*, 21–26.

Zhang, X., & Lu, X. (2013). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics*. doi:10.1093/applin/amt014.

Zhang, Y. (2005). Metacognition and college vocabulary teaching and learning. *Foreign Languages and Their Teching, 195*, 26–28.

Zhou, D. (2000). A Track investigation of English vocabulary of Chinese College students. *Foreign Language Teaching and Research, 32*(5), 356–361.

Zhu, C. (1994). *Psychology in Foreign Language teaching*. Shanghai: Shanghai Foreign Language Education Press.

Chapter 9
A Rasch-Based Approach for Comparison of English Listening Comprehension Between CET and GEPT

Quan Zhang, Guoxiong He and Huifeng Mu

Abstract This paper describes Rasch-based approach for comparison of English listening comprehension between two important English language tests, i.e., General English Proficiency Test (GEPT) in Taiwan and College English Test (CET) in China Mainland. A total of, 141 students of non-English majors of Jiaxing University were randomly chosen to take a mixed listening test with 15 questions from GEPT (high-intermediate) and 10 questions from CET (Band 4), respectively. All the data thus collected were processed using Gitest, with both subject ability and test difficulty compared. The results show that all the test items, as expected, are well moderated and calibrated. The correlation (= 0.61) shows that the difficulty level is fit for our subjects, but further confirmation is to be resorted by administering the same test items to the students of the similar ability level in Taiwan in the same manner. In the authors' view, this is a pioneer yet significant comparison. More cooperation and collaborative efforts are needed in the future.

Keywords English listening comprehension · Rasch model · College english test (CET) · General english proficiency test (GEPT)

9.1 Research Purpose

The motivation to compare these two tests remains for ages. Searching online shows that so far few significant researches (Gong 2002; Wu 2012) were ever conducted in this regard. Pacific Rim Objective Measurement Society (PROMS) 2013 held in Kaohsiung sparked the action and made such a research feasible. Also, the time has been matured to do such a comparison with focus on both test takers

Q. Zhang (✉)
College of Foreign Studies, University of Jiaxing, Jiaxing, People's Republic of China
e-mail: qzhang141@aliyun.com; proms2012_reg@hotmail.com

G. He · H. Mu
Institute of Language Testing, University of Jiaxing, Jiaxing, People's Republic of China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_9

ability and test item difficulty in terms of listening and reading comprehension in the Chinese context. Finally, Rasch model turns out to be the most appropriate approach to fulfill the task (Benjamin et al. 1979; Shichun 1985; Bachman and Palmer 1996; Trevor et al. 2007; Quan 2004, 2013).

In what follows is presented a Rasch-based approach recently conducted ad hoc for comparison of English listening and reading comprehension between GEPT in Taiwan and CET in China Mainland. The present paper would focus on listening, and the reading may resort to pp. 131–144.

9.2 Test Descriptions

GEPT in Taiwan is a test of English proficiency with five levels currently being administered: elementary, intermediate, high-intermediate, advanced, and quality, of which the High-Intermediate level is administered ad hoc for university students of non-English majors.[1]

CET Band-4 is a test of English proficiency for educational purpose designed by Shanghai Jiaotong University according to the requirements of college English teaching and administered only to sophomore students of non-English majors. CET has been administered ever since 1987 across China Mainland and even beyond. The count of CET test takers remains number one in today's world.[2]

At the first glance, GEPT and CET seem different. Academically, both are proficiency tests in nature. All that characterizes GEPT—multiple-choice question type, item analyses, test equating, and the use of scores—are found in CET. What is more, experts from both sides seek advice from the same testing authority, Professor Lyle. F. Bachman. Both tests are administered and used in Chinese-speaking context, and the test takers are Chinese speakers. English is their foreign language. Furthermore, even the tutorial manner and training or cramming classes run on and off campus setting on both sides are the same in nature, and the book markets are exactly the same.

9.3 Research Hypothesis

H_0 Both test items are well calibrated

H_0 Students in Chinese context are of the same ability.

[1] For more details, please visit http://www.gept.org.tw/.

[2] For more details, please visit http://www.cet.edu.cn/.

9.4 Research Method

To make sure our subjects give equal attention to the test items, both GEPT and CET listening and reading comprehension test items are mixed together in a test, and the test was administered as a compulsory middle-term test to our sophomore students of non-English major. To be more precise, it is the same group of subjects that tackled different test items within one test paper form in the view of language testing, the comparison or equating via common subjects in nature (Benjamin et al. 1979; Shichun 1985; Bachan and Palmer 1996).

9.5 Materials and Computer Program Used

For listening, we have six conversations with 15 questions from GEPT (Advanced) from Language Training and Testing (LTTC) (2015) and three passages with 10 questions from CET (Band 4), totaling to 25 questions.

For reading, we have three passages with 15 questions from GEPT (Advanced) and three passages with 15 questions from CET (Band 4), respectively, totaling to 30 questions. The present paper would focus on the listening. For more discussion and analyses of reading, interested readers may resort to PROMS 2014CN006 on pp. 131–144 of this book.

Rasch-based computer software Gitest 3[3]+ (1986v, 1989v) developed by Shichun (1985) and Wei (1989) was used to process all the data. In particular, Gitest 3+ was used to process all the data for 10 year MET[4] Equating Project sponsored by China Ministry of Education and has so far remained the one and only Rasch-based software developed by Chinese testing experts within China Mainland.

[3] For interested readers, please contact the author for details about GITEST. Ten features are listed below for reference:

1. Written in BASIC according to Rasch Model ;
2. It assumes binary (right-wrong) scoring;
3. Designed for applications of both CTT and Rasch to practical testing problems;
4. Maximum likelihood (ML);
5. Tests of fit for individual items;
6. Analysis of multiple subtests in one pass;
7. Item analysis and test paper evaluation and report;
8. Feedback for teaching and testing improvement
9. Linking of 2 test forms through common items (good for test equating);
10. 200 items/10,000 candidates/in a single run; (Benjamin 1979; Shichun 1985; Quan 2004, 2013).

[4] MET is abbreviated from Matriculation EnglishTest, a most competitive and influential entrance examination for higher education launched by Examination Authority under Ministry of Education, P.R.China. The annual participants amounts to 10 million or so.

9.6 Subjects

A total of 141 sophomore students of non-English majors from the Jiaxing University in Zhejiang Province were randomly selected to take the test. It is presumed that the sample is of intermediate level and homogenous in nature, best representative of students of non-English major students across China Mainland.

9.7 Data and Results

In this section are presented the tables obtained from Gitest 3+ indicating how well the two test items were calibrated and how the subject ability ranges.

Table 9.5 in Appendix I shows the item analysis for Item 1 of listening comprehension part. In total, there are 25 questions, so correspondingly there are 25 tables like this one for each item. Due to the limited space, the first table of Item 1 is presented here for discussion. For details, interested readers may refer to Appendix (Fig. 9.1).

Figure 9.2 in Appendix I shows both the difficulties of the reading item and the ability of the subject. The ability curve goes from −4 to +4 (logit) or shows all the possible scores, a typical Rasch curve. Another thing worth mentioning is that Items 16, 17, and 18 turn out to be very difficult for our students, and we are interested in knowing whether the same test items would be equally difficult for students in Taiwan.

9.8 Discussions and Conclusions

As shown in Tables 9.1, 9.2, 9.3, 9.4, 9.5, and 9.6 and Figs. 9.1 and 9.2 (Seen in Appendix I) based on the data processed by Gitest and addressed earlier, the discussion germane to listening can be summarized in at least three points as follows.

9.8.1 The listening comprehension test items of both sides are well moderated and calibrated with only four items (<0.3), and correlation coefficient is 0.61, showing higher correlation between the two tests. This is further confirmed by the factor analysis showing that both CET and GEPT listening belong to one factor. Such a finding is fully in conformity with the actual speaking environment on campus, where the language used for daily communication is in Chinese Mandarin, Hakka, or other dialects of the Chinese language; therefore, what is considered difficult to understand aurally on one side is also difficult to follow on the other.

9.8.2 Some similarities between the test items from the two tests are not only because they are moderated according to the similar framework standards of language proficiency but also because they are, in fact, designed for people sharing the same Chinese culture. This is, to a great extend, reinforcing the assumption proposed by Gong (2002) in their comparison between GEPT and PETS and by the test contents or the ideas inherent in the listening comprehension of both tests. This can be justified by the scores thus obtained by the students.

9.8.3 As few literature reviews could be available, further confirmation is needed by administering the same test items to the students of the homogenous background in Taiwan in the same manner. In a word, more analyses will be conducted in this regard.

9.9 Significances and Limitations

The present research, while focusing on the comparison of GEPT and CET, presents the research method via Rasch supported with real data analyses and thus can be concluded in at least two points as follows.

At the first place, the present study is the pioneer one ever conducted in comparison between CET and GEPT. Next, probably the most significant parts of the present research lie in the following:

(1) to show to our English teachers the importance of item analysis and test scoring with the help of Rasch;
(2) to demonstrate how item analysis and test scoring are actually conducted using Gitest 3+, the only Rasch-based software developed in 1980 s and still in use today; and
(3) how to understand the ideas regarding Rasch Model with detailed interpretation.

However, two corresponding limitations exist: small sample size and further justification needed to administer the same test items to the students of the homogenous background in Taiwan in the same manner.

Acknowledgments We would like to thank the anonymous reviewers and our foreign counterparts for suggestions that greatly helped improve the paper. Our thanks also go to Yang Hong of Jiaxing University who did one way or another to contribute to the present research project and their further cooperation will also be appreciated.

Appendix I

See Tables 9.1, 9.2, 9.3, 9.4, 9.5, and 9.6 and Figs. 9.1 and 9.2.

Table 9.1 GEPT and CET listening comprehension

```
•  P-VALUE AND R-BIS. CROSS TABLE
•  !!==========================================================================
•  !!R / P!!  0-.1   .1-.2   .2-.3   .3-.4   .4-.5   .5-.6   .6-.7   .7-.8   .8-.9   .9-1   !!TOTAL !!
•  !!===========================================================================!!
•  !! <.1 !!   0       0       0       0       0       0       0       0       0       0     !!   0   !!
•  !!.1-.2!!   0       0       0       1       0       0       1       0       0       0     !!   2   !!
•  !!.2-.3!!   0       0       0       0       1       1       0       0       0       0     !!   2   !!
•  !!.3-.4!!   0       0       3       1       1       0       0       0       0       0     !!   5   !!
•  !!.4-.5!!   0       0       2       1       2       1       2       0       0       0     !!   8   !!
•  !!.5-.6!!   0       0       0       2       1       4       0       0       0       0     !!   7   !!
•  !!.6-.7!!   0       0       0       0       1       0       0       0       0       0     !!   1   !!
•  !!.7-.8!!   0       0       0       0       0       0       0       0       0       0     !!   0   !!
•  !!.8-.9!!   0       0       0       0       0       0       0       0       0       0     !!   0   !!
•  !!.9-1 !!   0       0       0       0       0       0       0       0       0       0     !!   0   !!
•  !!==========================================================================!!
•  !!TOTAL!! 0       0       5       5       6       6       3       0       0       0     !!  25   !!
•  !!==========================================================================!!
•  !!        !!   VD 5%   !      D 15%     !       I 60%         !  E 15%    !     VE 5%        !! TOTAL !!
•  !!        !!     2     !        3       !        15                3       !        2       !!   25   !!
•  !!============================================================================!!
•  !!TOTAL!!       0     !          5       !        20          !    0      !        0        !!   25   !!
•  !!==========================================================================!!
•  ! PASS    !!    0     !          5       !        16          !    0      !        0        !!  21    !!
•  !!============================================================================!!
```

where *P* in each row indicates probability (difficulty) and *R* in each column, the discrimination index in bi-serial. Difficulty level is expressed respectively in VD, i.e. Very difficult: (<0.1), only below 0.1 probability of getting the correct answer; *D*, i.e. Difficult: (= 0.1 ~ 0.3) only between 0.1 and 0.3 probability of getting the correct answer, *I*, i.e. Intermediate: (0.3 ~ 0.7), showing the probability of getting the correct answer ranges from 0.3 to 0.7, *E*, i.e. Easy: E (0.7 ~ 0.9), showing the probability of getting the correct answer between 0.7 and 0.9, and VE, i.e. Very easy: (>0.9) showing the chance of getting the correct answer being above 0.9, and where TOTAL indicates the total number of items calibrated and in the row indicated by PASS shows the number of calibrated items that are believed to be fit for the group of samples (See Benjamin et al. 1979; Shichun 1985; Quan 2004, 2013).

Table 9.2 GEPT and CET listening comprehension

- Lis TEST TABLE
- TOTAL NO. OF ITEMS: 25 TOTAL NO. OF SUBJECTS: 143

```
DATE OF TEST:        DATE OF ANALYSIS: 08-02-2014
------------------------------------------------------------------------------------------------
 Mean !   SD   ! Varn.!  p+   !   pd   !   R11   ! Rbis ! aVALUE ! Skew.!   Kurt. !
------------------------------------------------------------------------------------------------
 11.05    3.19   10.18   0.44    13.58    0.46     0.44    0.38     0.11     -0.19
------------------------------------------------------------------------------------------------
 GEPT (Scr)      CET (Scr)
   6.85             4.2
------------------------------------------------------------------------------------------------
GEPT (logit)     CET (logit)
 -0.06393         0.0802
------------------------------------------------------------------------------------------------
```

where Mean refers to the mean scores of the whole examinees; SD, the standard deviations of the whole examinees; Varn. the variants based on the whole examinees; *P*+ probability of correct answers; Pd Δ value, difficulty parameter based on probability; *R*11 by Kuder-Richardson20, reliability, this value should be over 0.9; aVALUE reliability parameter, also called α value by Cronbach formular, this value should be over 0.8; Rbis discrimination index (in the unit of bi-serial); Skewness score distribution value, of which 0 indicating normal distribution; above 0, indicating positive skewness, showing the test items more difficulty; below 0, indicating negative skewness, showing the test items easier; Kurtosis score distribution height, of which 0 indicating normal; above 0 showing "narrower", i.e. small range between the scores; below 0, indicating "flat", i.e. big range between scores; Difficulty VD (<0.1), D (= 0.1 ~ 0.3), I (0.3 ~ 0.7), E (0.7 ~ 0.9), VE ($>$0.9) (See Benjamin et al. 1979; Shichun 1985; Quan 2004, 2013)

Table 9.3 GEPT and CET listening comprehension

- **Lis SUBTEST TABLE GEPT**
-
- NO. OF ITEMS: 15 (FROM 1 TO 15)

```
------------------------------------------------------------------------------------------------------
 Mean   !  SD    ! Varn. !  p+    !  pd     !  R11   !  Rbis !  Skew. !  Kurt. !
  6.85     2.25     5.08    0.46    13.43     0.36      0.41     0.09     -0.51
------------------------------------------------------------------------------------------------------
Difficulty ! TOTAL!   NO.(<.3) !     ITEMS
-----------!---------.!-------------!-----------------------------------------------------------------
 VD    !    0    !       0      !
   D   !     4   !       0      !
   I   !    11   !       3      !    3 10 11
   E   !     0   !       0      !
 VE    !     0  !       0      !
----------!----------!-------------!--------------------------------------
```

As indicated above, no very difficult items, no easy items and no very easy items calibrated in this part. There are four difficult items and 11 intermediate items, but the discrimination index of the latter three items were found below 0.3. They are specified as Item 3, Item 10 and Item 11

Table 9.4 GEPT and CET listening comprehension

```
              Lis SUBTEST TABLE                CET

      NO. OF ITEMS:    10    (FROM    16    TO    25    )
      --------------------------------------------------------------------------------------------------------
       Mean    !   SD      ! Varn. !   p+      !   pd       !   R11     !   Rbis !   Skew. !    Kurt. !
       4.20        1.78        3.17        0.42        13.81        0.30        0.47      -0.06          -0.67
      --------------------------------------------------------------------------------------------------------
      Difficulty ! TOTAL ! NO.(<.3) !       ITEMS
      ------------!-----------!-----------!------------------------------------------------------------------------
         VD     !     0     !     0     !
          D     !       1   !       0     !
          I      !       9   !     1      !   23
          E       !     0   !       0     !
         VE     !     0     !       0   !
      -----------!-----------!------------!------------------------------------------------------------------------
```

where item 23 was calibrated as difficult item yet whose discrimination index was found below 0.3. All the other items are fit for the group

Table 9.5 Lis item analysis

```
 DATE OF TEST:                                           DATE OF ANALYSIS:08-02-2014
 !-----------!----------!--------------!--------!------!-----!-----!-------!-------!-------!---------!-------- --!--------!-----!--------!-----!--------!---
 ! TEST      CODE !   ITEM NO. !      Pt      !      Pi    !      Pd      !        P       !      MA      !    MB !    MC     !     MD    ! MO       !
 ! Lis       ! GEPT              1          !    13.58     13.43     15.30          0.28          12.76       11.19     12.80      14.95      0.85    !
 !-----------!-----------!--------!----------!---------!---------!--------!---------!--------!--------!--------!------ --!--------!-----!--------!-----!-------
 ! NO.OF CAN.!        KEY    !    Ar      !      Br      !      Cr      !      Dr      !      A       !       B       !      C      !      D     !      O      !
 !    141            !        D       !     0.05         0.36         0.03       0.41          39              35              27              40            0        !
 !-----------!----------!-------!----------!---------!---------!--------!---------!------- !---------!------!----- --!--------!-----!--------!-----!--------!----
```

where MA, MB, MC, and MD and MO in the first row refer, respectively, to the scores or means obtained by the test takers who chose A, B, C, and D and who did not tick any choice. Such a score is transformed into 13 as mean and 4 as standard deviation (SD), indicating 13 as equal to the average of the whole population and greater than 13 indicating above the average and smaller than 13 indicating lower than the average. Here MO shows the test takers whose scores are far below the average (See Benjamin et al. 1979; Shichun 1985; Quan 2004, 2013)

Table 9.6 Item analysis: listening comprehension GPET and CET

ITEM DIFFICULTIES FROM 141 PERSONS

ALL POSSIBLE SCORES ON THE TEST

Item No.	Difficulty	Stand Error(di)	Corr. No.	Score(br)	Stand Error(br)
1	0.725	0.193	1	-3.348	1.028
2	0.878	0.199	2	-2.597	0.747
3	-1.086	0.187	3	-2.132	0.627
4	-0.688	0.178	4	-1.783	0.559
5	0.839	0.197	5	-1.496	0.514
6	-0.949	0.183	6	-1.247	0.484
7	0.276	0.180	7	-1.023	0.462
8	0.580	0.188	8	-0.814	0.445
9	0.762	0.194	9	-0.621	0.434
10	-0.255	0.175	10	-0.433	0.426
11	-0.316	0.175	11	-0.254	0.421
12	-0.657	0.178	12	-0.084	0.419
13	-0.657	0.178	13	0.084	0.419
14	0.309	0.181	14	0.251	0.421
15	-0.720	0.179	15	0.421	0.425
16	0.342	0.182	16	0.601	0.433
17	0.148	0.178	17	0.784	0.443
18	-0.132	0.175	18	0.984	0.458
19	0.148	0.178	19	1.194	0.478
20	1.043	0.207	20	1.427	0.505
21	-0.688	0.178	21	1.688	0.543
22	-0.439	0.176	22	2.010	0.601
23	0.616	0.189	23	2.456	0.706
24	0.085				

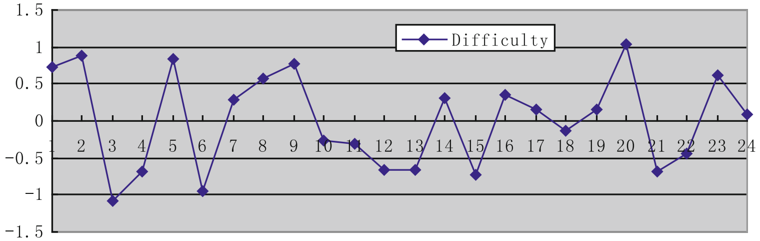


Fig. 9.1 Item difficulty of GEPT and CET listening comprehension

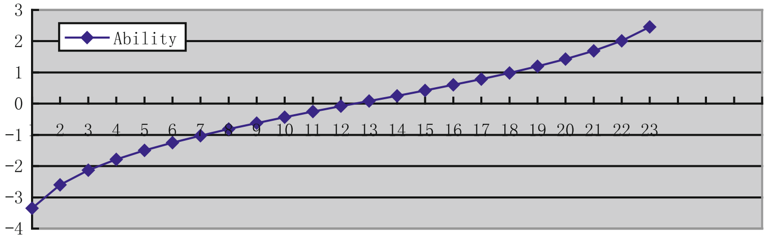


Fig. 9.2 Ability of GEPT and CET listening comprehension

Appendix II

```
 [1] "                                    Lis ITEM ANALYSIS"
 [2] ""
 [3] "DATE OF TEST:                                    DATE OF ANALYSIS:08-02-2014"
 [4] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!-!-----!-----!-----!-----!-!-----!-----!-----!-----!---!"
 [5] "! TEST CODE ! ITEM NO.   ! Pt    !   Pi    !   Pd    !     P    !   MA    !    MB    !    MC    ! MD    ! MO    !"
 [6] "! Lis             ! GEPT 1       !13.58    13.43    15.30    0.28      12.76      11.19      12.80    14.95    0.85    !"
 [7] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!-------!-----!-----!-----!-------!-----!-----!---------!"
 [8] "! NO.OF CAN.!       KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
 [9] "!      141          !        D         ! 0.05    0.36    0.03    0.41        39        35        27        40         0 !"
[10] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!--------!-----!-----!-----!-------!-----!---!-----!---!"

[12] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[13] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[14] "! Lis             ! GEPT 2      !13.58 13.43 15.64    0.26 12.15 13.11 15.28 10.87    0.85 !"
[15] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[16] "! NO.OF CAN.!       KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[17] "!    141            !     C           ! 0.20 -0.02    0.45    0.34        51        34        36        20         0 !"
[18] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
```

```
[20] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[21] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[22] "! Lis              ! GEPT 3      !13.58 13.43 11.12    0.68 13.39 12.20 13.09 11.67    0.85 !"
[23] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[24] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[25] "!    141            !      A           ! 0.19    0.12 -0.01    0.21        96        15        10        20         0 !"

[28] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[29] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[30] "! Lis              ! GEPT 4      !13.58 13.43 12.03    0.60 10.04 11.75 14.26 10.46    0.85 !"
[31] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[32] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[33] "!    141            !      C           ! 0.39    0.24    0.48    0.35        11        34        84        12         0 !"
[34] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[36] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[37] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[38] "! Lis              ! GEPT 5      !13.58 13.43 15.55    0.26 14.85 12.34 10.86 12.73    0.85 !"
[39] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[40] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[41] "!    141            !      A           ! 0.37    0.11    0.32    0.08        37        21        17        66         0 !"
[42] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[44] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[45] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[46] "! Lis              ! GEPT 6      !13.58 13.43 11.43    0.65 12.08 10.60 13.94 11.49    0.85 !"
[47] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[48] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[49] "!    141            !      C           ! 0.14    0.41    0.41    0.16        18        26        92         5         0 !"
[50] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[54] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[55] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[56] "! Lis              ! GEPT 7      !13.58 13.43 14.27    0.38 12.82 11.90 11.32 14.50    0.85 !"
```

```
[57] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[58] "! NO.OF CAN.!     KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O     !"
[59] "!    141          !     D          ! 0.03    0.22    0.27    0.37       32       35       21       53         0 !"
[60] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[61] ""
[62] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[63] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[64] "! Lis             ! GEPT 8      !13.58 13.43 14.96    0.31 12.08 15.40 10.70 13.65    0.85 !"
[65] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[66] "! NO.OF CAN.!     KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O     !"
[67] "!    141          !     B          ! 0.10    0.53    0.58 -0.13         6       44       54       37         0 !"
[68] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[69] ""
[70] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[71] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[72] "! Lis             ! GEPT 9      !13.58 13.43 15.38    0.28 12.00 13.09 14.54 11.95    0.85 !"
[73] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[74] "! NO.OF CAN.!     KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O     !"
[75] "!    141          !     C          ! 0.20 -0.02    0.32    0.18       35       40       39       27         0 !"
[76] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[77] ""
[78] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[79] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[80] "! Lis             ! GEPT 10    !13.58 13.43 13.04    0.50 12.82 11.92 12.28 13.85    0.85 !"
[81] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[82] "! NO.OF CAN.!     KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O     !"
[83] "!    141          !     D          ! 0.02    0.23    0.11    0.26       12       41       18       70         0 !"
[84] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[85] ""
[86] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[87] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[88] "! Lis             ! GEPT 11    !13.58 13.43 12.89    0.51 12.17 13.03    9.72 13.76    0.85 !"
[89] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
```

```
 [90] "! NO.OF CAN.!       KEY     ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
 [91] "!    141            !     D          ! 0.17 -0.01    0.39    0.24       39        23         7        72          0 !"
 [92] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
 [93] ""
 [94] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
 [95] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
 [96] "! Lis             ! GEPT 12    !13.58 13.43 12.10    0.59 13.64 10.16 14.33 10.60    0.85 !"
 [97] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
 [98] "! NO.OF CAN.!       KEY     ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
 [99] "!    141            !     C          !-0.09    0.53    0.50    0.33       14        32        83        12          0 !"
[100] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"


[104] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[105] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[106] "! Lis             ! GEPT 13    !13.58 13.43 12.10    0.59 11.01 14.40 11.26 10.54    0.85 !"
[107] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[108] "! NO.OF CAN.!       KEY     ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[109] "!    141            !     B          ! 0.32    0.53    0.29    0.34       22        83        23        13          0 !"
[110] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!---------!-----!-----!-----!-------!-----!---!-----!---"
[111] ""
[112] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[113] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[114] "! Lis             ! GEPT 14    !13.58 13.43 14.34    0.37 15.41 11.19 10.81 12.77    0.85 !"
[115] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[116] "! NO.OF CAN.!       KEY     ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[117] "!    141            !     A          ! 0.59    0.28    0.45    0.04       52        18        39        32          0 !"
[118] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[119] ""
[120] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[121] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[122] "! Lis             ! GEPT 15    !13.58 13.43 11.96    0.60 10.21 11.27 14.10 13.26    0.85 !"
[123] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[124] "! NO.OF CAN.!       KEY     ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[125] "!    141            !     C          ! 0.48    0.26    0.43 -0.04       25        16        85        15          0 !"
```

```
[126] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[127] ""
[128] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[129] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[130] "! Lis              !CET 16        !13.58 13.81 14.42    0.36 15.06 11.51 12.32 11.67    3.57 !"
[131] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[132] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[133] "!    141            !      A           ! 0.50    0.27    0.12    0.25        51        28        29        33          0 !"
[134] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[135] ""
[136] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[137] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[138] "! Lis              !CET 17        !13.58 13.81 13.97    0.40 11.37 15.55 11.03 11.43    3.57 !"
[139] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[140] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[141] "!    141            !      B           ! 0.25    0.66    0.37    0.31        17        57        31        36          0 !"
[142] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[143] ""
[144] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[145] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[146] "! Lis              !CET 18        !13.58 13.81 13.32    0.47 11.81 11.77 14.70    9.90    3.57 !"
[147] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[148] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[149] "!    141            !      C           ! 0.19    0.27    0.50    0.41        21        43        66        11          0 !"
[150] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[154] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[155] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"
[156] "! Lis              !CET 19        !13.58 13.81 13.97    0.40 10.38 11.99 13.49 14.64    8.06 !"
[157] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[158] "! NO.OF CAN.!        KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"
[159] "!    141            !      D           ! 0.51    0.14 -0.10    0.43        34        12        36        57          2 !"
[160] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[161] ""
```

```
[162] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[163] "! TEST CODE ! ITEM NO. ! Pt   ! Pi   ! Pd   !   P   ! MA   ! MB   ! MC   ! MD   ! MO    !"
[164] "! Lis         !CET 20      !13.58 13.81 16.00   0.23 12.47 14.87 12.55 12.50   5.82 !"
[165] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[166] "! NO.OF CAN.!      KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O    !"
[167] "!    141         !      B         ! 0.09    0.35    0.09    0.11        27        32        40        41         1 !"
[168] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[169] ""
[170] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[171] "! TEST CODE ! ITEM NO. ! Pt   ! Pi   ! Pd   !   P   ! MA   ! MB   ! MC   ! MD   ! MO    !"
[172] "! Lis         !CET 21      !13.58 13.81 12.03   0.60 14.53 12.74   9.99 10.40   3.57 !"
[173] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[174] "! NO.OF CAN.!      KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O    !"
[175] "!    141         !      A         ! 0.59    0.04    0.48    0.44        84        12        21        24         0 !"
[176] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[177] ""
[178] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[179] "! TEST CODE ! ITEM NO. ! Pt   ! Pi   ! Pd   !   P   ! MA   ! MB   ! MC   ! MD   ! MO    !"
[180] "! Lis         !CET 22      !13.58 13.81 12.61   0.54    8.81 11.97 11.96 14.50   3.57 !"
[181] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[182] "! NO.OF CAN.!      KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O    !"
[183] "!    141         !      D         ! 0.61    0.19    0.16    0.51        15        31        19        76         0 !"
[184] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[185] ""
[186] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[187] "! TEST CODE ! ITEM NO. ! Pt   ! Pi   ! Pd   !   P   ! MA   ! MB   ! MC   ! MD   ! MO    !"
[188] "! Lis         !CET 23      !13.58 13.81 15.05   0.30 13.90 12.30 13.76 12.46   3.57 !"
[189] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[190] "! NO.OF CAN.!      KEY     ! Ar   ! Br   ! Cr   ! Dr   !   A   !   B   !   C   !   D   !   O    !"
[191] "!    141         !      C         !-0.14    0.17    0.16    0.09        20        53        43        25         0 !"
[192] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[193] ""
[194] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
[195] "! TEST CODE ! ITEM NO. ! Pt   ! Pi   ! Pd   !   P   ! MA   ! MB   ! MC   ! MD   ! MO    !"
```

```
[196] "! Lis          !CET 24       !13.58 13.81 13.82    0.42 14.42 11.13 12.42 12.83    3.57 !"

[197] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[198] "! NO.OF CAN.!       KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"

[199] "!    141           !     A           ! 0.38    0.36    0.11    0.02        59        33        33        16          0 !"

[200] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"


[204] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[205] "! TEST CODE ! ITEM NO. ! Pt    ! Pi    ! Pd    !    P    ! MA    ! MB    ! MC    ! MD    ! MO      !"

[206] "! Lis          !CET 25       !13.58 13.81 13.25    0.48 11.06 14.83 11.37 11.99    8.06 !"

[207] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"

[208] "! NO.OF CAN.!       KEY      ! Ar    ! Br    ! Cr    ! Dr    !    A    !    B    !    C    !    D    !    O      !"

[209] "!    141           !     B           ! 0.30    0.55    0.33    0.15        18        67        38        16          2 !"

[210] "!-----------!----------!-----!-----!-----!-----!-----!-----!-----!-----!------!---!-----!-----!-----!-------!-----!---!-----!---"
```

(Benjamin et al. 1979; Shichun 1985)

References

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

Benjamin, D., Wright & Mark, H., Stone. (1979). Best Test Design Rasch Measurement. MESA

Language Training and Testing Center (LTTC). (2015). The general english proficiency test: Level descriptors. [Online] available: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm (March 6, 2015)

Shichun, G. (1985). *Standardized examination: theory practice and method*. China: Guangdong Higher Education Press.

Trevor, G., Bond & Christine, M., Fox. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd Edition) Lawrence Erlbaum. ISBN-13: 978-0805842524 ISBN-10: 0805842527

Quan, Z. (2004). *Item analysis and test equating for language testing in practice*. Beijing: Higher Education Press.

Quan, Z. (2013). A pilot study based on Rasch into the appropriateness of the TOEIC bridge test for Chinese students: Status quo and prospect. In Q. Zhang & H. Yang (Eds.), *Pacific rim objective measurement symposium 2012 conference proceeding*. Germany: Springer.

Gong, Y. (2002). Comparative studies of GEPT and PETS in Taiwan and China, Selected Papers from the Eleventh International Symposium on English Teaching/Fourth Pan-Asian Conference.

Wei, L. (1989). *MET (1985−1987) and Data Analyses*. Guangdong Education Press. ISBN7-5406-0595-2/G.594

Wu, Mei. (2012). Comparing PETS and GEPT in China and Taiwan. *English Language Teaching, 5*(6), 48–51.

Chapter 10
A Rasch-Based Approach for Comparison of English Reading Comprehension Between CET and GEPT

Hong Yang and Mingzhu Miao

Abstract A Rasch-based approach for comparison of English reading comprehension between two important English language tests, i.e., The General English Proficiency Test (GEPT) in Taiwan and College English Test (CET) in China mainland. A total of 141 students of non-English majors of Jiaxing University were chosen to take a reading comprehension test with mixed test items from both GEPT and CET. There were three passages with 15 questions from GEPT (High-intermediate) and three passages with 15 questions from CET (Band 4), respectively, totalling to 30 questions. All the data were collected and processed using GiTest 3+, comparing both student ability and test difficulty. The results show that the test has a mean score of 16.39, with 55 % of the answers correct, indicating it is a medium-level test, and the standard deviation is 3.66 against the expected standard deviation of 4.48, having a range of 17, indicating that the scores are well distributed. On the whole, the distribution of the scores is negatively skewed. The standard error of measurement is ±2.34. The reading comprehension test items of both examinations are well moderated and thus well calibrated.

Keywords English reading comprehension · Rasch model · CET · GEPT

10.1 Background

College English Test (CET) Band-4 is a national test of English as a foreign language in China mainland, with the purpose of examining the English proficiency of sophomore students of non-English majors and ensuring that Chinese undergraduates reach the required English levels specified in the *National College English Teaching Syllabuses* (NCETS), covering four language skills of listening,

H. Yang (✉) · M. Miao
Faculty of Foreign Languages, Jiaxing University, Jiaxing, Zhejiang, China
e-mail: redbeetle2008@sina.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_10

reading, writing, and speaking. CET has been administered for nearly 30 years since 1987, and CET certificate is of special significance to college students for employment purposes across China Mainland.

General English Proficiency Test (GEPT), a test of English language proficiency, is developed and administered by the Language Training and Testing Center (LTTC) in Taiwan since 1999, and the development project came to completion in July 2002. The five levels of the test are currently being administered: elementary, intermediate, high-intermediate, advanced, and superior, covering the four language skills of listening, reading, writing, and speaking in a balanced way, with the goal of improving the general English proficiency level of English learners in Taiwan. GEPT has been regarded as the benchmark for selecting talents, recruiting new students, or evaluating by public and private institutions.

10.2 Research Purpose

Due to enormous common characteristics found in the two tests, this has been evoking the researchers' great interest in seeking comparison and contrast between the tests. Although researches on GEPT-related and CET-related topics have been undertaken, respectively, to get a good understanding of the reliability and validity of the GEPT and CET as well as their relationship with other English language tests, few significant studies have been ever conducted in comparing the GEPT and CET. It is motivated by this that our research group explores tentatively both the test takers' ability and item's difficulty using one test with mixed test items for listening and reading comprehension from CET and the GEPT (High-intermediate level) in Jiaxing, Zhejiang Province, China Mainland. And the present paper would deal with the reading, and the listening will be discussed in a separate paper.

10.3 Literature Review

Reading, one of the four language skills, is so important that almost all the large-scale language tests include reading part at home and abroad, such as CET4, CET6, PETS, TEM4, TEM8, TOEFL, and GRE, with a larger proportion in each whole paper. In recent years, it has been universally accepted worldwide that language learners' reading ability can be elicited and compared from their language performance in reading tests in a standard manner and under uniform conditions. People are more concerned about such issues related to reading test as how to define the construct "reading ability" (construct validity), what to be tested (content validity), what kind of forms to be used (face validity), and how to evaluate the relevance of two individual and reliable tests (criterion-related validity), and so on.

10.3.1 Construct Validity

According to Bachman (1999), in defining a construct, the test developer "needs to make a conscious and deliberate choice to specify particular components[1] of the ability or abilities to be measured in a way that is appropriate to a particular situation" and will most likely "base the construct definition on the specific components of language ability that are included in the course syllabus" or "on the components described in a theory of language ability." Bachman's work laid the theoretical foundation for the research on language ability. And According to Heaton (2000), "reading comprehension, in which questions are set to test the students' ability to understand the gist of a text and to extract key information on specific points in the text." And Heaton identified a series of specific skills involving reading ability. Many large-scale proficiency tests follow these doctrines at home and abroad.

Li and Zeng (2011) studied the construct validity based on Bachman's Communicative Language Ability by exploring the relationship between the use of cognitive strategies and meta-cognitive strategies and reading comprehension. Various data collected from the studies based on test takers' reading scores or reading strategies, question types, observation, or introspection were used to validate some reading tasks (Alderson 1990; Jin and Wu 1998; Zou et al. 2002; Kong 2011; Shi 2015).

10.3.2 Content Validity

Content validity refers to what it superficially appears to measure. The content of reading tests should include extensive subject matters and various styles, i.e., representativeness of materials selected and sampling adequacy. CET4 covers the materials related to biography, society, culture, everyday knowledge, scientific knowledge, and so on. And the general requirement is that the test takers can read and understand the main idea and the supporting facts and details. GEPT (High-intermediate) focuses more on the topics related to daily life experience and work with local characteristics. The requirements[2] are test takers who pass this level have a generally effective command of English and can handle a broader range of topics and can read different types of articles on concrete and abstract topics. Both of the tests use authentic materials in line with "real-life" approach (Bachman 1999).

[1] Some language developers have different views from Bachman on language ability, regarding it as a holistic, unitary ability.

[2] Language Training and Testing Center (LTTC). (2012). The General English Proficiency Test: Level Descriptors. https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/hi_intermediate.htm.

10.3.3 Face Validity

Face validity means a test is going to measure what it is supposed to measure, in other words, surface credibility or public acceptability. In most cases, reading tests are direct tests in which test takers are required to complete some reading tasks after reading some sentences or some articles. The arguments are mainly on the effects of the question types on test takers (Liu 1998; Alderson 1983) and the patterns of relationship among specific question types, reading strategies, and reading performance (Zou et al. 2002). It seems that people more favor alternative assessment because any question type has its own advantages and disadvantages. The reading test in IELTS has been affirmed by many experts in testing field because of its diversity, authenticity, and positive wash-back. (Li et al. 2014)

10.3.4 Comparison of Two Tests

Wu (2012) compared the test levels, the test contents, and the scoring weight for test sections of GEPT and PETS on the whole and tentatively concluded "generally speaking, the GEPT reflects a theory of test design based on the Communicative Approach to English acquisition, while the PETS reflects a lingering effect of the traditional grammar teaching." Cai (2006) also thinks that CET hasn't emphasized the measurement of language use ability, with validity and credibility not balanced. Gong (2002) concluded that "there is a real possibility of comparability between GEPT and PETS." Is there a possibility of comparability between GEPT and CET4 in reading part? Will the test takers in China mainland or in Taiwan show the same performance in the two tests?

According to Li and Luo (2014), CTT has some limitations. There are some differences between CTT and Rasch Model:

> CTT methods generate ordinal variables instead of interval variables, but the latter is often a prerequisite for conducting parametric studies. These shortcomings make CTT less effective in both educational and psychological measurement. Compared to CTT, the Rasch model generates objective measures which are both sample- and instrument-free (Bond and Fox 2007). It produces genuine interval variables and put the person abilities and item difficulties along one logit scale.

Therefore, they think that Rasch Model has an advantage over CTT in constructing and testing the psychometric properties of a measuring instrument. Engelhard (2013) make a comparison of the two research traditions (test-score and scaling). He thinks that the test-score tradition emphasizes test scores and determines the sum of item responses with the measurement error and sources of error variance in measurement quantified, while the scaling tradition focuses on seeking invariant measurements. In his book, he takes the following assertion seriously:

"If invariant measurement is achieved, then it is possible to simultaneously locate both items and persons on variable maps." Some Rasch-based studies have been done on testing concurrent validity (Zhang 2013).

There are very few studies on comparison between CET and GEPT. However, all the above-mentioned studies have provided the robust grounds for comparing validity and reliability of the two tests. This study aims to make a Rasch-based comparison of both student ability and test difficulty in terms of reading comprehension using mixed test items of the two tests in randomly selected 141 college students from Jiaxing University.

10.4 Research Design

10.4.1 Research Questions

The reading part in GEPT (High-intermediate) is a three-section test with 45 multiple-choice questions, sentence completion, cloze, and reading comprehension (multiple-choice items), and there are two sections in CET4 with reading in-depth (two short passages with multiple-choice items and one short passage with passage completion) and so-called "fast reading" (with one long passage with matching items).

Similarly, in the reading part of CET4 and GEPT (high-intermediate), questions testing the students' reading skills are included such as understanding explicitly stated information, understanding relations within the sentence, generalizing and drawing conclusion, skinning and scanning, and so on. Using some questions in the form of multiple-choice items chosen from reading comprehension in GEPT (High-intermediate) and reading in-depth in CET4, this study tries to answer the following two questions:

(1) How well are the two test items calibrated?
(2) Do college students in China mainland show the same ability in the reading comprehension of the two tests?

10.4.2 Sample

A total of 141 students of non-English majors of Jiaxing University were randomly chosen as subjects who had been learning English almost for 1 year at college. And they hadn't known anything about GEPT before they took the mixed test.

10.5 Method

Three passages with 15 questions from GEPT (High-intermediate) and three passages with 15 questions from CET (Band 4) were chosen, respectively. There were a total of 30 questions, among which 21 items were concerning understanding details and 9 needing making inferences. In order to make sure that the subjects would give equal attention to the test items, all the six passages were mixed and the test was administered to the subjects (non-English majors) as a compulsory mid-term test. The subjects must carefully weigh up the four choices for each question and select the best answer after reading each passage. They were allowed to finish the test within 50 min.

10.5.1 Data Collection and Analysis

After the test, the subjects' answer sheets were scanned with binary (right–wrong) scoring (1 marks for correct responses and 0 marks for incorrect responses). All the data collected were processed using the Rasch-based computer software Gitest 3+ (1986v, 1989v) developed by Gui Shichun and Li Wei (See Footnote 1), including item analysis, correlation, test scores, and comparison. Gitest 3+ was designed for applications of both CTT and Rasch to practical testing problems. The results presented subsequently are typically Rasch based.

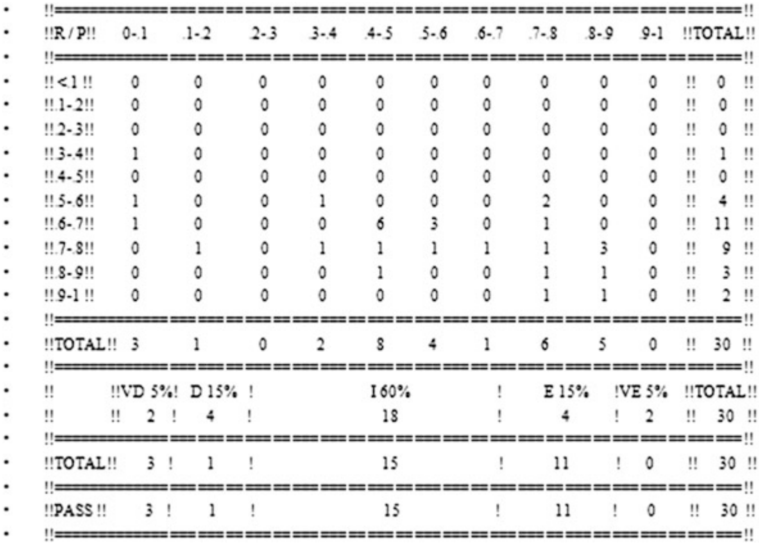
10.6 Results

In the following, tables (or diagram) such as *P* value and *R*-BIS, cross table, reading comprehension test table, reading comprehension subset tables as well as item analysis, reading comprehension GPET and CET, are used to show the results obtained from Gitest 3+ analysis.

In Table 10.1, *P* in each row indicates probability (difficulty) and *R* in each column indicates the discrimination index in biserially. Difficulty level is expressed, respectively, as VD, i.e., Very difficult: (<0.1); *D*, i.e., Difficult: (= 0.1 $\sim$ 0.3); *I*, i.e., Intermediate: (0.3 $\sim$ 0.7); *E*, i.e., Easy: *E* (0.7 $\sim$ 0.9); and VE, i.e., Very easy: (>0.9). From Table 10.1, we can see that 2 items are VD, 4 items are *D*, 18 items are *I*, 4 items are *E*, and 2 items are VE. TOTAL indicates the 30 items are all well calibrated and the row PASS shows all the items are believed to be fit for the group of samples.

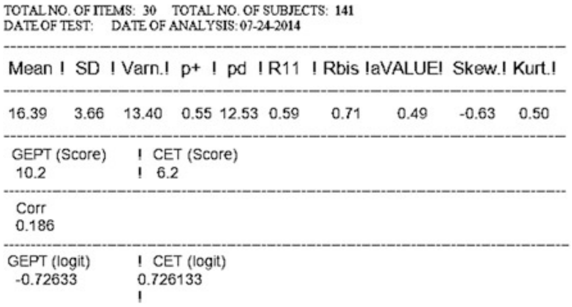
From Table 10.2, we can see that the test has a mean score of 16.39, *P*+ (probability of correct answers) 0.55, indicating it is a medium-level test. The standard deviation is 3.66 against the expected standard deviation 4.48, having a range of 17, indicating the scores are well distributed. On the whole, the distribution of the scores is negatively

Table 10.1 *P*-value and *R*-BIS. cross table



```
•  !!==================================================================!!
•  !!R / P!!   0-.1   .1-2   .2-3   .3-.4   .4-.5   .5-.6   .6-.7   .7-.8   .8-.9   .9-1   !!TOTAL!!
•  !!==================================================================!!
•  !! <1 !!     0      0      0      0      0      0      0      0      0      0    !!   0  !!
•  !!.1-2!!     0      0      0      0      0      0      0      0      0      0    !!   0  !!
•  !!.2-3!!     0      0      0      0      0      0      0      0      0      0    !!   0  !!
•  !!.3-.4!!    1      0      0      0      0      0      0      0      0      0    !!   1  !!
•  !!.4-.5!!    0      0      0      0      0      0      0      0      0      0    !!   0  !!
•  !!.5-.6!!    1      0      0      1      0      0      0      2      0      0    !!   4  !!
•  !!.6-.7!!    1      0      0      0      6      3      0      1      0      0    !!  11  !!
•  !!.7-.8!!    0      1      0      1      1      1      1      1      3      0    !!   9  !!
•  !!.8-.9!!    0      0      0      0      1      0      0      1      1      0    !!   3  !!
•  !!.9-1 !!    0      0      0      0      0      0      0      1      1      0    !!   2  !!
•  !!==================================================================!!
•  !!TOTAL!!    3      1      0      2      8      4      1      6      5      0    !!  30  !!
•  !!==================================================================!!
•  !!        !!VD 5%!  D 15%  !              I 60%           !     E 15%    !VE 5%   !!TOTAL!!
•  !!        !!   2  !    4   !                18            !       4      !   2    !!  30  !!
•  !!==================================================================!!
•  !!TOTAL!!      3  !    1   !                15            !      11      !   0    !!  30  !!
•  !!==================================================================!!
•  !!PASS !!      3  !    1   !                15            !      11      !   0    !!  30 !!
•  !!==================================================================!!
```

Table 10.2 Reading comprehension test table



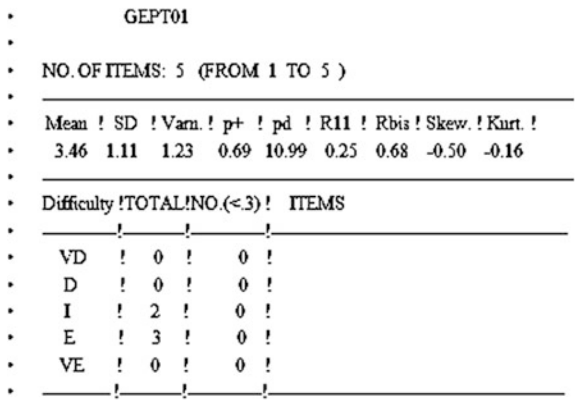
```
TOTAL NO. OF ITEMS: 30    TOTAL NO. OF SUBJECTS: 141
DATE OF TEST:    DATE OF ANALYSIS: 07-24-2014
----------------------------------------------------------------------
Mean !  SD  ! Varn.!  p+  !  pd  ! R11  ! Rbis !aVALUE!  Skew.!  Kurt.!
----------------------------------------------------------------------
16.39    3.66   13.40   0.55  12.53  0.59    0.71     0.49    -0.63    0.50
----------------------------------------------------------------------
 GEPT (Score)      !  CET (Score)
 10.2              !  6.2
----------------------------------------------------------------------
 Corr
 0.186
----------------------------------------------------------------------
GEPT (logit)       !  CET (logit)
 -0.72633          0.726133
                   !
```

skewed. The standard error of measurement is ±2.34. Kurtosis is 0.5, showing "narrower", i.e., small range between the scores.

Analysis of each component in Table 10.3 shows that GEPT01 has a mean score of 3.46, having 69.15 % of answers correct, indicating it is a medium-level component. On the whole, in this component GEPT01, there is 0 items that does not meet the requirement.

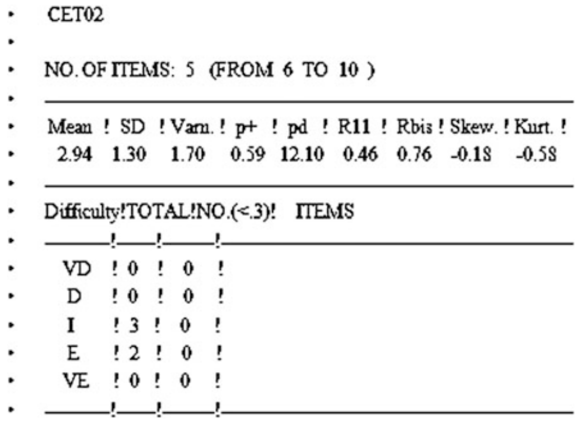
Analysis of each component in Table 10.4 shows that CET02 has a mean score of 2.94, having 59 % of answers correct, indicating it is a medium-level component. On the whole, in this component CET02, there is 0 items that does not meet the requirement.

Table 10.3 Reading subtest tables: GEPT01



```
•             GEPT01
•
•   NO. OF ITEMS:  5  (FROM  1  TO  5 )
•   ________________________________________________________________
•   Mean  !  SD  ! Varn. !  p+  !  pd  !  R11  !  Rbis ! Skew. ! Kurt. !
•    3.46   1.11    1.23    0.69   10.99   0.25    0.68   -0.50   -0.16
•   ________________________________________________________________
•   Difficulty !TOTAL!NO.(<.3) !   ITEMS
•   ________!_______!________!______________________________
•     VD     !   0   !     0   !
•      D     !   0   !     0   !
•      I     !   2   !     0   !
•      E     !   3   !     0   !
•     VE     !   0   !     0   !
•   ________!_______!________!______________________________
```

Table 10.4 Reading subtest tables: CET02



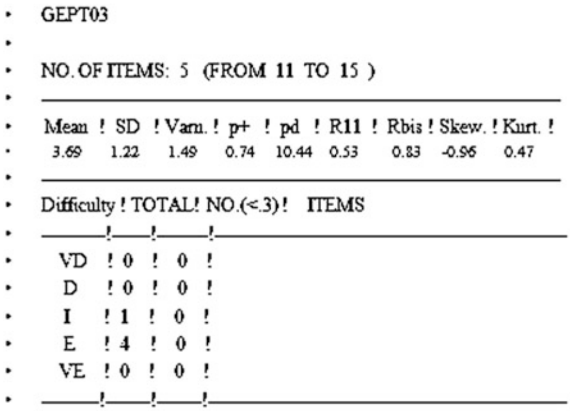
```
•   CET02
•
•   NO. OF ITEMS:  5  (FROM  6  TO  10 )
•   ________________________________________________________________
•   Mean  !  SD  ! Varn. !  p+  !  pd  !  R11  !  Rbis ! Skew. ! Kurt. !
•    2.94   1.30    1.70    0.59   12.10   0.46    0.76   -0.18    -0.58
•   ________________________________________________________________
•   Difficulty!TOTAL!NO.(<.3)!   ITEMS
•   _______!____!_____!______________________________________
•     VD    ! 0  !  0   !
•      D    ! 0  !  0   !
•      I    ! 3  !  0   !
•      E    ! 2  !  0   !
•     VE    ! 0  !  0   !
•   _______!____!_____!______________________________________
```

Analysis of each component in Table 10.5 shows that GEPT03 has a mean score of 3.69, having 74 % of answers correct, indicating it is a fairly easy component. On the whole, in this component GEPT03, there is 0 items that does not meet the requirement.

Analysis of each component in Table 10.6 shows that CET04 has a mean score of 1.05, having 21 % of answers correct, indicating it is a difficult component. On the whole, in this component CET04, there is 0 items that does not meet the requirement.

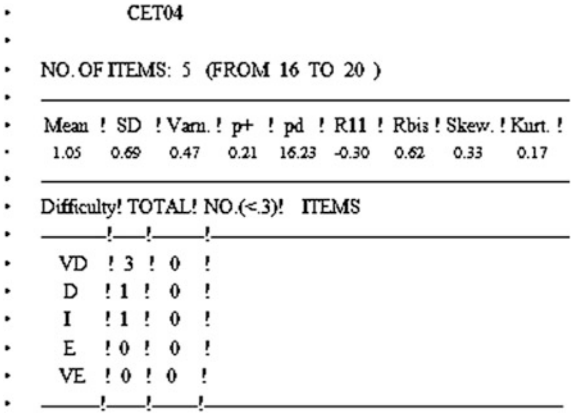
Analysis of each component in Table 10.7 shows that GEPT05 has a mean score of 3.01, having 60 % of answers correct, indicating it is a medium-level component. On the whole, in this component GEPT05, there is 0 items that does not meet the requirement.

Table 10.5 Reading subtest tables: GEPT03



```
•   GEPT03
•
•   NO. OF ITEMS: 5 (FROM 11 TO 15 )
•   ____________________________________________________________
•   Mean ! SD ! Varn. ! p+ ! pd ! R11 ! Rbis ! Skew. ! Kurt. !
•   3.69   1.22   1.49   0.74   10.44  0.53   0.83   -0.96   0.47
•   ____________________________________________________________
•   Difficulty ! TOTAL! NO.(<.3)!  ITEMS
•   _______!____!_____!______________________________
•     VD   ! 0  !  0  !
•     D    ! 0  !  0  !
•     I    ! 1  !  0  !
•     E    ! 4  !  0  !
•     VE   ! 0  !  0  !
•   _______!____!_____!______________________________
```

Table 10.6 Reading subtest tables: CET04



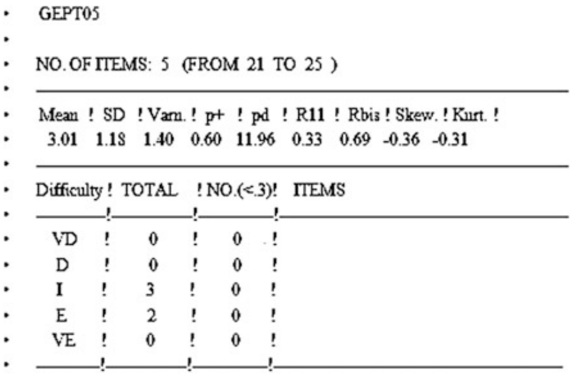
```
•             CET04
•
•   NO. OF ITEMS: 5 (FROM 16 TO 20 )
•   ____________________________________________________________
•   Mean ! SD ! Varn. ! p+ ! pd ! R11 ! Rbis ! Skew. ! Kurt. !
•   1.05   0.69   0.47   0.21   16.23  -0.30   0.62   0.33   0.17
•   ____________________________________________________________
•   Difficulty! TOTAL! NO.(<.3)!  ITEMS
•   _______!____!_____!______________________________
•     VD   ! 3 ! 0   !
•     D    ! 1 ! 0   !
•     I    ! 1 ! 0   !
•     E    ! 0 ! 0   !
•     VE   ! 0 ! 0   !
•   _______!____!_____!______________________________
```

Analysis of each component in Table 10.8 shows that CET06 has a mean score of 2.24, having 45 % of answers correct, indicating it is a medium-level component. On the whole, in this component CET06, there is 0 items that does not meet the requirement.

From Fig. 10.1, we can see the 30 item difficulties (logit) from 141 test takers based on Rasch Model, with 1 parameter logistic normal metric ($D = 1.7$). All the 30 items are calibrated in the same scale with five levels of difficulties (VD, *D*, *I*, *E*, and VE) well identified.

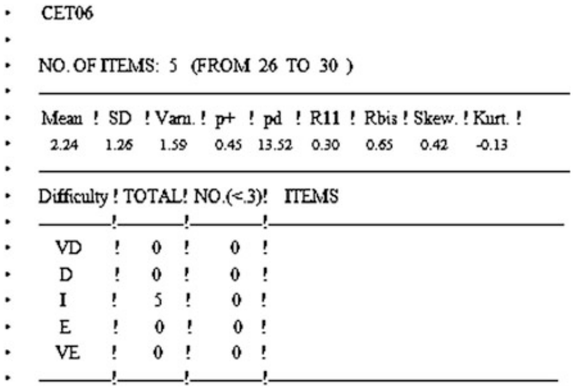
Figure 10.2 shows both the reading item difficulties and the subject ability. The ability curve goes from −4 to +4 (logit), or shows all the possible scores, a typical

Table 10.7 Reading subtest tables: GEPT05



```
•  GEPT05
•
•  NO. OF ITEMS:  5  (FROM  21  TO  25 )
•  ______________________________________________________________
•  Mean  !  SD  ! Varn. !  p+  !  pd  !  R11  !  Rbis ! Skew. ! Kurt. !
•   3.01   1.18   1.40   0.60   11.96   0.33   0.69  -0.36  -0.31
•  ______________________________________________________________
•  Difficulty !  TOTAL    ! NO.(<.3)!   ITEMS
•  _________!__________!_________!__________________________
•     VD    !     0     !     0    !
•     D     !     0     !     0    !
•     I     !     3     !     0    !
•     E     !     2     !     0    !
•     VE    !     0     !     0    !
•  _________!__________!_________!__________________________
```

Table 10.8 Reading subtest tables: CET06



```
•  CET06
•
•  NO. OF ITEMS:  5  (FROM  26  TO  30 )
•  ______________________________________________________________
•  Mean  !  SD  ! Varn. !  p+  !  pd  !  R11  !  Rbis ! Skew. ! Kurt. !
•   2.24    1.26    1.59    0.45   13.52   0.30    0.65    0.42    -0.13
•  ______________________________________________________________
•  Difficulty ! TOTAL! NO.(<.3)!   ITEMS
•  _________!________!_________!__________________________
•     VD     !   0   !     0    !
•     D      !   0   !     0    !
•     I      !   5   !     0    !
•     E      !   0   !     0    !
•     VE     !   0   !     0    !
•  _________!________!_________!__________________________
```

Rasch curve. Another thing worth mentioning is that Items 16, 17, and 18 turn out to be very difficult for our students, and we are interested in knowing whether the same test items would be equally difficult for students in Taiwan.

10.7 Discussion and Conclusions

Based on the data presented earlier, we've reached to three conclusions as follows:

5.1 Item analysis shows that there are 5 % very easy items, 15 % easy items, 60 % medium-level items, 15 % difficult items, and 5 % very difficult items. On the whole, the test has good discrimination power. There are 0 items that do not

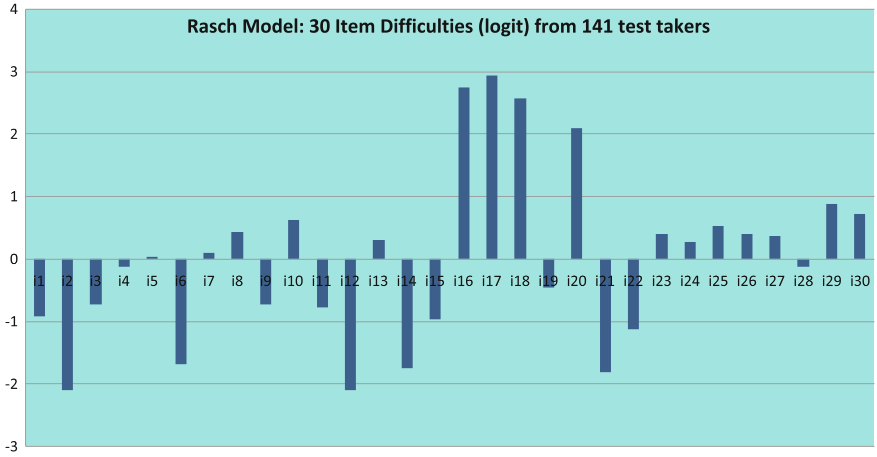


Fig. 10.1 GPET and CET reading comprehension

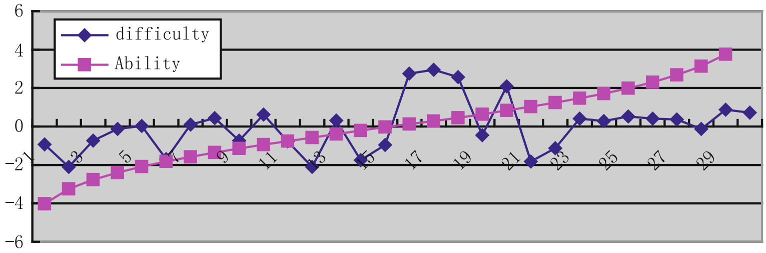


Fig. 10.2 Reading Item difficulty and person ability

meet the requirement. Factor analysis of the test shows that it accounts for 30.23 % of the knowledge and skills of the examinees. It also shows that GEP01 and GEP05 belong to one factor. This is extremely important and need to be further confirmed by administering the same test items in Taiwan.

5.2 Our data analysis of the mixed test mentioned earlier shows that there exists possibility of comparability between CET4 and GEPT (High-intermediate) on both student ability and item difficulty. There are some similarities between the two tests because they share the similar framework standards of language proficiency (reading comprehension). As Gong (2002) points out that “a comparative study [between the PETS and the GEPT] can be especially important not only because language testing is a central issue in foreign-language teaching, but also because GEPT and PETS are practically designed for people sharing a very similar Chinese culture.”

5.3 Test contents or the ideas inherent in the reading comprehension of both examinations are well chosen. This can be justified by the scores thus obtained by the students. Reading comprehension questions on both sides are well designed.

The reading comprehension test items of both examinations are well moderated and thus well calibrated. This is by no means easy, showing both examinations are appropriate for students of non-English major in China mainland, but we need to further confirm it by administering the same test items to the students of the similar ability level in Taiwan in the same manner. More analyses will be conducted.

Acknowledgments We would like to thank the anonymous reviewers and our foreign counterparts for suggestions that greatly helped improve the paper. Our thanks also go to Prof. Zhang Quan of Jiaxing University who organized such a research project and contributed both academically and financially to make such a research feasible and successful.

Appendix

See Tables 10.9 and 10.10.

Table 10.9 Item difficulties from 141 persons

Item no.	Difficulty	Stand error (di)
1	−0.926	0.199
2	−2.107	0.28
3	−0.734	0.192
4	−0.125	0.178
5	0.032	0.176
6	−1.694	0.244
7	0.094	0.176
8	0.434	0.176
9	−0.734	0.192
10	0.622	0.178
11	−0.771	0.193
12	−2.107	0.28
13	0.31	0.176
14	−1.756	0.249
15	−0.966	0.201
16	2.749	0.307
17	2.952	0.333
18	2.575	0.287
19	−0.451	0.184
20	2.095	0.243
21	−1.82	0.254
22	−1.135	0.209
23	0.403	0.176
24	0.279	0.176
25	0.528	0.177

(continued)

Table 10.9 (continued)

Item no.	Difficulty	Stand error (di)
26	0.403	0.176
27	0.372	0.176
28	−0.125	0.178
29	0.882	0.183
30	0.718	0.18

Table 10.10 All possible scores on the test

Corr. No.	Score (br)	Stand error (br)
1	−4.027	1.038
2	−3.254	0.761
3	−2.767	0.644
4	−2.396	0.577
5	−2.089	0.533
6	−1.821	0.502
7	−1.58	0.479
8	−1.358	0.462
9	−1.146	0.448
10	−0.948	0.438
11	−0.759	0.431
12	−0.572	0.425
13	−0.394	0.422
14	−0.216	0.42
15	−0.034	0.42
16	0.128	0.422
17	0.285	0.425
18	0.456	0.43
19	0.637	0.438
20	0.827	0.447
21	1.028	0.46
22	1.24	0.476
23	1.467	0.496
24	1.713	0.52
25	1.994	0.552
26	2.308	0.592
27	2.682	0.649
28	3.133	0.732
29	3.751	0.889

References

Alderson, C. (1983). The cloze procedure and proficiency in english as a foreign language. In J. W. Oller Jr (Ed.), *Issues in language testing research*. Newsbury House: Rowley.

Alderson, J. C. (1990). Testing reading comprehension skills (part two) getting students to talk about taking a reading test (a pilot study). *Reading in a Foreign Language, 7*, 465–503.

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

Bachman, L. F. (1999). *Language testing in practice*. Shanghai: Shanghai Foreign Language Education Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erblaum. ISBN-13: 978-0805842524 and ISBN-10: 0805842527

Cai, J. G. (2006). Exploring the reform of CET in China from the perspective of the reform of STEP in Japan. *Foreign Language Teaching Abroad, 1*, 41–56.

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

Gong, Y. (2002). *Comparative Studies of GEPT and PETS in Taiwan and China*. Selected papers from the eleventh international symposium on english teaching/fourth pan-asian conference.

Heaton, J. B. (2000). *Writing english language Tests*. Beijing: Foreign Language Teaching and Research Press.

Jin, Y., & Wu, J. (1998). Investigating the validity of reading test in CET through introspection. *Foreign Language World, 2*, 47–52.

Kong, W. (2011). Validating the TEM4 reading tasks from the verbal report data collected from students' think-aloud protocol. *Foreign Language Testing and Teaching, 3*, 1–13.

Language Training and Testing Center (LTTC). (2015). The general english proficiency test: Level descriptors. [Online] available: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm (March 6, 2015)

Li, G. M., & Zeng, Y. Q. (2011). A study of construct validity of communicative language ability. *Modern Foreign Languages, 4*, 389–396.

Li, L. R., & Luo, S. Q. (2014). Development and validation of a test measuring language analytic ability for chinese fl learners: A Rasch-based pilot study. *Chinese Journal of Applied Linguistics, 4*, 194–212.

Li, X., Xiu, X. D., & Su, F. Y. (2014). A comparative study of cognitive validity of question types in the NEMT and IELTS reading tests. *Ludong University Journal, 5*, 64–68.

Liu, J. D. (1998). The effects of various test methods on the test takers' performance in reading testing. *Foreign Language Teaching and Research, 2*, 48–52.

Shi, Y. L. (2015). Investigating construct validity of long reading-matching item in CET4 through analyzing the test takers' think-aloud verbal reports. *Foreign Language Testing and Teaching, 1*, 24–31.

Wu, Mei. (2012). Comparing PETS and GEPT in China and Taiwan. *English Language Teaching, 5*(6), 48–51.

Zhang, Q. (2013). A pilot study based on Rasch into the appropriateness of the TOEIC bridge test for Chinese students: Status quo and prospect. In Q. Zhang & H. Yang (Eds.), *Pacific rim objective measurement symposium 2012 conference proceeding*. Germany: Springer.

Zou, S., Zhang, Y. L., & Zhou, Y. M. (2002). A study of potential relationships among question types, test takers' reading strategies and their reading test scores—An attempt to further validate TEM4 reading tests. *Foreign Languages and Their Teaching, 5*, 19–22.

Chapter 11
Application of Rasch Model in Test Paper Quality Analysis—With the Case of Reading Literacy Pilot Test Paper for Grade 5–6 Students in Guangxi

Jing Gong and Dehong Luo

Abstract The methods of applying Rasch Model in the analysis of test paper quality are as follows: Wright Map which shows the reader a general knowledge of the whole survey, Multidimensionality Investigations which is used to examine whether test paper measures the latent trait (reading dimension of subjects), Item Fit Order, Bubble Diagram, and so on. Analysis result shows that the reading literacy pilot test paper for Grades 5–6 students in Guangxi meets the requirement generally. Specifically speaking, with covering different ability students and making up reasonably difficulty items, the test items reach the expected test effect. For better and higher application efficiency, the suggestion for usage is that the choice of Rasch Model function and result analysis methods be based on different test goals.

Keywords Rasch model · Reading literacy · Pilot test · Test quality

Chinese government realizes more than ever the importance of assessing students' ability scientifically. The *Implementation Opinions in Deepening Reform on Examination and Recruitment System* issued recently by Chinese State Council emphasizes the fairness, scientificity, and objectivity of choosing talented person, which sets higher requirements for content and measurement method of test. Scientific measurement is an important standard of examining students' ability, teaching objectives, and advancing quality of education. In the recent years, Rasch Model has been gradually applied in the fields of education, healthcare, and psychology in the world, but it is still rather new for most teachers in primary education, especially in China. They seldom use this tool to improve students' ability or scientize test quality. Our thesis takes the *Reading Literacy Pilot Test Paper for Guangxi Grade 5–6 Grade Students* as an example for elaborating how to apply Rasch Model in the quality analysis of the test paper.

J. Gong · D. Luo (✉)
Educational College of Guangxi University, Guangxi, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_11

11.1 Rasch Model Instruction

Created by Rasch, the Danish mathematician and educationalist in 1960, Rasch Model is a probabilistic model for measuring the latent trait according to the responses to the item made by responders (Rasch 1960). Rasch believes that reaction probability of subjects to specific item can be formulated by simple functions: $\mathrm{Log}_e(P_{ni1}/(1-P_{ni1})) = B_n - D_i$ (John 2012). In Rasch dichotomous model, responsively correct and incorrect items are denoted separately by 1 and 0. B_n represents the ability level of person N, D_i represents the difficulty of item I. Probability P_{ni1} is that person n of ability B_n scores 1 on item i of difficulty D_i. $1-P_{ni1}$,P_{ni0}, is the probability of scoring 0. On the left side of the equation we have log(odds), which provides the units for the right side of the equation, so B_n and D_{i} are measured in “log-odds units.” All of these give the Rasch Model an important peculiarity: measurers can measure the item difficulty and person ability at the same time with the Rasch Dimension; therefore, item difficulty and person ability are both independent of each other and compared with each other. This peculiarity has a wide range of applications in the American grading measurement research. According to the Rasch Model, the famous Metametrics Inc. exploited the grading measurement software Lexile Analyzer, which is applied to calculate students’ reading ability and text difficulty by particular Lexile Points as the basis of individual education implementation. Having been widely used in different kinds of schools in 50 states of America, it reports Lexile Points for the text difficulty of 30 million books and for reading ability of half of American Students. (Luo and Yu 2013).

11.2 Research Method

11.2.1 Reading Literacy Pilot Test of Guangxi Grade 5–6 Students

For examining students’ literacy of reading for interest and application, reading literacy is classified into three kinds according to the psychological process of human cognition in this pilot test, which are, from low to high and from simple to complex, literacy of access and extract, of integration and interpretation, and of introspection and evaluation. Based on text category classification of different perspectives, five text categories, literature text, continuous text, incontinuous text, informational text, and comprehensive test, have been chosen, and two responsive question formats which are Subjective Constructive Question (SCQ) and Objective Constructive Question (OCQ) have been designed. Specific distribution is shown as in Table 11.1 (C stands for OCQ, while S for SCQ).

Table 11.1 Question formats distribution

	Reading articles				
Reading literacy	Article 1	Article 2	Article 3	Article 4	Article 5
Access and extract	S2	C6	S8	S12	S16, S17, C18
Integration and interpretation	S1, C3	S5	S9	C10, S11, C14, S15	C19
Introspection and evaluation	S4	S7		S13	

11.2.2 Research Objects

There were about 30,000 students who joined the self-designed reading literacy pilot test in Guangxi in 2013, among which there were 12,000 Grade 5–6 students. The internationally and widely used means of dichotomous stratified random sampling is used, thus ensuring the representativeness of the population.

As the first step, five cities in Guangxi are chosen, and schools in which there are Grade 5–6 students are divided into three types: provincial capital city schools, other city schools and country schools, from which 41 schools with the random sampling method have at last been sampled. Also, in order to ascertain the representativeness of the schools in each type, the sample schools have intentionally covered and balanced typical ones of different qualities from low to high.

As the second step, in the 41 schools, we randomly sample the data of 6912 students from 153 classes for analysis.

11.2.3 Analytical Method

Bond&Foxsteps1.0.0 software and Rasch Model are used to analyze the data.[1] This software can be compatible with EXCEL and SPSS. For meeting the basic requirements of Rasch Model operation, after data are imported, we used dichotomous scoring for OCQ and multiple scoring for SCQ. For OCQ method, we applied 0-1 form, 0 for fault answers and 1 for correct answers, and for the SCQ method, we applied 0-1 or 0-1-2 forms. These different scoring methods are based on the Category Probability Curves, which resulted from the analysis of Bond&Foxsteps1.0.0 software.

[1] Free download site of Rasch Model is http://www.winsteps.com/bigsteps.htm. 3000 variates and 20,000 subjects' data could be applied by the free version. This site provides free and prepaid workshop and thesis online and offline.

11.3 Applying Rasch Model in the Quality Analysis of Test Paper

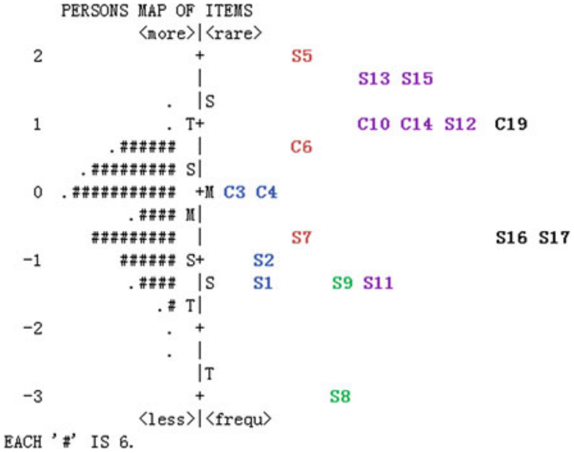
Applying Rasch Model in the quality analysis of test is taken is done as follows: first of all, Wright Map shows the reader a general knowledge of the whole survey; second, Multidimensionality Investigations examines whether test paper measures the latent trait (reading dimension of subjects) or not; and then Item Fit Order and Bubble Diagram, which measures the quality of test.

11.3.1 Wright Map

At the beginning of analyzing the quality of test, in order to know the whole difficulty of test, researchers usually use the Wright Map to analyze the data. Wright Map can intuitively present the corresponding relation between item difficulty and person ability in one dimension as shown below (Fig. 11.1).

Midline is the Logit ruler, and it is the important medium of comparing the item difficulty with person ability. M is the acronym of Mean and represents the average level. S is the acronym of One Standard Error, meaning one standard deviation away from the M. T is the acronym of Two Standard Errors, meaning two standard deviations away from the Mean.

On the right-hand side is the distribution of 18 items of pilot test; the 18th item was deleted because of the misprint. In the rest of the 18 items, on the right side from up to down, the item difficulty decreases progressively. The most difficult item is the S5, and the easiest one is S8; five articles are distributed from left to right in the proper order, and the concrete item number of each article is shown in Fig. 11.1. The total lowest difficulty and the highest difficulty among the five articles are separately *Understanding the Subway Line Map* and the literature text, while the



```
      PERSONS MAP OF ITEMS
              <more>|<rare>
   2                +         S5
                    |               S13 S15
                 .  |S
   1             . T+               C10 C14 S12  C19
           .######  |         C6
        .######### S|
   0  .###########  +M C3 C4
             .#### M|
         #########  |         S7                 S16 S17
  -1        ###### S+      S2
             .####  |S     S1     S9 S11
                .# T|
  -2             .  +
                 .  |
                    |T
  -3                +             S8
              <less>|<frequ>
EACH '#' IS 6.
```

Fig. 11.1 Wright Map

rest of three articles have reasonable medium difficulty. On the left-hand side is the distribution of person ability, and symbol # represents a certain number of students. From up to down, the students reading abilities decrease progressively.

The Wright Map shows that the item difficulty is designed reasonably. As we can see, the difficulty distribution of test is about 5 logit, and the students' ability distribution is about 3.5 logit. The mean of item difficulty and students' reading ability are approximated with each other, which means that the difficulty distribution of test covers all the students' reading ability. It can be concluded that the reading literacy pilot test compiled by the research group can measure the subjects' reading ability precisely.

11.3.2 Multidimensionality Investigations

After knowing the whole difficulty of the test, researchers would adopt the Multidimensionality Investigations to analyze how many dimensions are there in the test. Because the *Reading Literacy Pilot Test for Guangxi Grade 5–6 Students* is aimed at assessing the students' reading ability, the dimension should be limited to the students' reading ability, being devoid of any other dimensions. If it is so, it can be justified that the students' reading ability can be examined by all the items of this test, which conforms to the whole test objective. Otherwise, researchers should turn to a further analysis to find out problem items to amend and perfect.

The main analysis method of Multidimensionality Investigations is the Principle Component Analysis of Residuals, PCRA.

If the minimum eigenvalue is less than 3 in the residual models, it demonstrates that the data are in accordance with one dimension. In the Fig. 11.2, 1, 2, 3, 4, and 5 represent five residual factors; the eigenvalue of the first factor is 2.4, less than 3; the eigenvalues of the other four factors are separately 1.7, 1.4, 1.3, and 1.3. Clearly, all of these factors are less than 3. It illustrates that the *Reading Literacy Pilot Test for Guangxi Grade 5-6 Students* is in accordance with one-dimensional test of Rasch Model. All items of this test are sharing only one Rasch dimension, justifying the test of students' reading literacy ability.

11.3.3 Item Fit Order and Bubble Chart

If Wright Map and Multidimensionality Investigations place extra emphasis on the overall analysis of the situation, Item Fit Order and Bubble Chart pay attention to estimating the total quality of the test from the independent item.

Table 11.2 is the fit index statistics of items, including the measure of average score of item difficulty, standard error, infit and outfit mean square (MNSQ), and correlation coefficient. In the test of goodness of item fit, whether the value of infit and outfit MNSQ is in the range of 0.5–1.5 is an important indicator for estimating

Fig. 11.2 Variance component scree plot

```
       +--+--+--+--+--+--+--+--+--+
 100%+     T                                +
       |                                    |
V 63%+                                      +
A       |       M                            |
R 40%+              U                       +
I      |                                   |
A 25%+                                      +
N       |                                   |
C 16%+                                      +
E      |                                    |
   10%+                                      +
L      |                                    |
O    6%+                 1                   +
G       |                                   |
|   4%+                     2                +
S      |                        3   4   5   |
C   3%+                                      +
A       |                                   |
L   2%+                                     +
E      |                                    |
D    1%+                                     +
       |                                    |
 0.5%+                                      +
       +--+--+--+--+--+--+--+--+--+
           TV MV UV U1 U2 U3 U4 U5
          VARIANCE COMPONENTS
```

Table 11.2 Item fit order

Item	Measure	S.E.	Infit MNSQ	Outfit MNSQ	Corr.
S1	−1.94	0.1	1.66	1.62	0.39
C3	−0.17	0.09	0.22	0.23	0.17
C4	0.01	0.1	0.34	0.35	0.19
S16	−0.68	0.1	2.67	2.68	0.08

whether the item fits well with Rasch Model or not (Zi 2010). If the value of Infit and outfit MNSQ is less than 0.5, then it means that the item belongs to the overfit item. Overfit item indicates that all the high-level students answer the item right and all the low-level students answer the item wrong, which is so ideal that it is difficult to achieve. If the value of infit and outfit MNSQ is more than 1.5, then it means that the item belongs to the underfit item. Underfit item indicates that all the high-level students answer the item wrong, and all the low-level students answer the item right, which is against the truth of human cognition and its function.

Among all the items of test, the value of items 1, 3, 4, and 16 are not in the range of 0.5–1.5, which manifests that the four items don’t fit the Rasch Model well enough. Items 1 and 16 are the underfit items, and the other two, Items 3 and 4, are the overfit items. What’s more, the standard error value of each item is about 0.1, and such tiny value demonstrates that this test can estimate students’ ability precisely and has a high reliability. Correlation coefficients of all items are in the range from 0 to 1, which means every item has the same goal as that of the test.

Bubble Chart not only depicts the fitting index vividly but also shows the size of standard error, which helps researcher to confirm the latent items that don’t fit Rasch Model well. Radius of bubble reflects standard error, and the smaller the bubble size and standard error, the more precise the result of the test measurement would be. Vertical axis represents the measurement value of item difficulty, while horizontal axis is the value of outfit MNSQ (if the value is in the range of 0.5–1.5, it would be considered to fit Rasch Model), which helps the researcher get a visual knowledge of all the item difficulties inside and outside value range of outfit MNSQ.

In Fig. 11.3, all the items are lined out. The standard error of Item 8 is the biggest, which means it can’t measure reading ability of the most part of students precisely. Except Items 1, 3, 4, and 16, the rest of items are in the range of 0.5–1.5, meaning fitting the Rasch Model well. This result is the same as that of fit item order (Table 11.2).

In a word, according to the analysis of Item Fit Order and Bubble Chart, items 1, 3, 4, and 16 are the unfit items and Item 8 has the biggest standard error. So, researcher needs further analysis of these items and then decides the revision plans.

Item 8 is the easiest one in the test. From the Wright Map, we can see there are a minority of subjects to be at the same level of this item, which makes a larger standard error. But it is normal to cover all levels of item difficulty in a piece of test paper, so we need to reserve item 8.

Items 3 and 4 are overfit items which are the ideal outcomes of our expectation. The interference of this kind of items is small in the data analysis. Although the result is too ideal to believe, it is suggested to reserve them. In the next test,

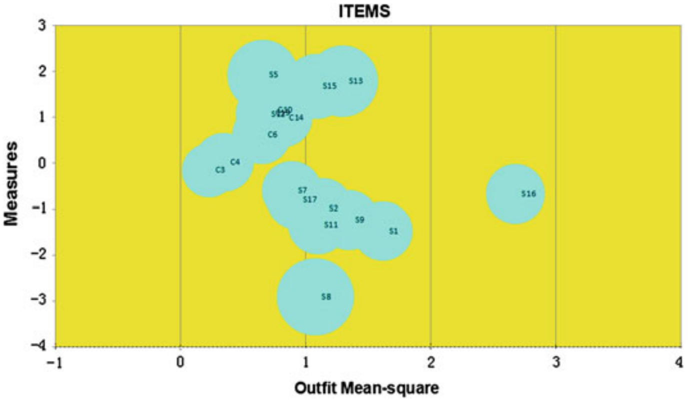


Fig. 11.3 Bubble Chart

researcher needs to compare and contrast the two analysis outcomes with each other for further analysis.

Items 1 and 16 are the underfit items. Having observed the students' answering test papers, we came to sum up the following results. For Item 1, most subjects with high reading ability aren't attentive enough in analyzing the question. They carelessly mistake *the amount of fireworks and crackers of New Year's eve* for *the amount of fireworks and crackers of the fifth day of lunar year.* Although they make the sense of *the amount of fireworks and crackers of the fifth day of lunar year* and also have the ability to solve this kind of problem, they can't get scores. The question item itself is reasonable and needs reserving. For Item 16, it is related to life experience, but there are some indigestible words in it. Students with the experience of cooking dishes can solve this item according to their life experience, although they don't necessarily have the high reading ability. However, for those high reading ability students, but without similar life experience, if they can't understand these indigestible words, they would not make it yet. Therefore, we need to amend it by clarifying the ambiguity of certain words and shielding its empirical information.

In conclusion, this pilot test is a high-quality test with a reasonable difficulty that covers all different ability levels of subjects and with the exception of Item 16 needing modifying, most of the items reach our expected test result.

11.4 Some Suggestions for Better Application of Rasch Model in Test Paper Analysis

Based on *Reading Literacy Pilot Test of Guangxi Grade 5–6 Students*, the aforementioned content includes the steps of analyzing test paper quality with Rasch model, selection of measurement index and result explanation separately. Different adoption of Rasch model functions, selection of its index, and explanation of its result will result in different analysis, so the next content will be focused on the some tips for better and flexible application.

11.4.1 Selecting Functions of Rasch Model Based on the Measurement Goal

Rasch Model is an ability measurement model, so except for the reading literacy test mentioned earlier, it is appropriate for measurement of all the subjects, such as Chinese, Mathematics, English, Physics, and so on. But because each subject has its own test goal, researcher must select analysis functions of Rasch Model according to the practical issue. For example, a math teacher wants to know whether there are differences in geometry achievement ability between girls and boys, instead of adopting the aforementioned methods, he or she should use the Differential Item Function to test his hypothesis.

11.4.2 Selecting and Explaining Measurement Index of Rasch Model Based on the Measurement Goal

This thesis examines the reading ability of subjects, so in Multidimensionality Investigations, all the items should conform to one-dimensional test that demonstrates the consistency between the test of item characteristics and the test objective. That is to say, the selection of index would be changed with the test objective changing. For example, there are four components of ability in English including listening, speaking, reading, and writing, and for measurement, there are four dimensions. If Multidimensionality Investigation is taken, we couldn't use the analysis index as mentioned earlier mechanically. Now researchers need to prove whether this English test has four dimensions through the Multidimensionality Investigations and whether these four dimensions point to listening, speaking, reading and writing. It means that the whole test should be conforming to the Multidimensionality Investigations. However, if we aim at analyzing each kind of items solely, they should conform to the one-dimensional test.

This pilot test is designed to have some acquaintance of the current situation of reading literacy of Guangxi Grade 5–6 students, so we try to make it sure that the test average difficulty is close to the students' average reading ability and that the difficulty distributions of items cover different reading ability of different subjects so that students' reading ability are tested precisely. Therefore, the important indicators of examining the quality of test are whether students' reading ability is comparable to difficulty in test or not, and whether difficulty in distributions cover all different levels of subjects or not. But if the goal of test paper is to determine the gap between students' ability and teaching objectives, researchers should compile items according to the teaching objectives. Only in this way could the gap between overall level of students and goal of examination be found out and could the gap between the ability of every student and goal of examination be detected. What's more, Rasch Model can also be used to locate that a particular student can't make it in a particular item and a particular class haven't accommodated and assimilated some particular knowledge.

11.4.3 Explaining and Disposing Unfit Items Based on the Measurement Goal

If unfit items are found out in the analysis of Item Fit Order and Bubble Chart, is it inevitable to amend or delete them? The answer should be that *it depends on*. Rasch Model is just a tool which is used to find out items that possibly affect test quality. It helps researchers to detect there are some problems with some items, which is its mission. A further decision and action should be based on the analysis of the content and students' responsive information. If all the items conform to the requirement of examined goal and in the zone of proximal development of students, and there are no empirical items which are independent of test content, even if they don't pass through fitting test, it can also be reserved for further usage and investigation. Therefore, a scientific test tool is concerned, it can't be regarded as a scientific one until several rounds have been taken to analyze, revise, verify, and redesign. Only in this way, can we apply it in tracing developmental level of students' reading literacy and predicting their developmental trend.

Acknowledgments Funded by the following foundations: Chinese National Science Foundation *Measuring Scores of Levels of Reading Comprehension Ability and Text Difficulty*, Chinese Educational Ministry Foundation (10JZD022) and Guangxi University Key Subject Groups of 211 Project *Cooperation and Development of Sino-ASEAN Economy and Trade*. Acknowledgement of Guangxi Jieli Publishing Company and Guangxi Educational Science Institute for data collection assistance.

References

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Institute of Educational Research.

John, M. L. (2012). A user's guide to winsteps [EB/OL]. winsteps.com.

Luo, D., & Yu, J. (2013). America Lexile level reading framework: Individual reading instruction and measurement tool. *Modern Primary Education, 10*, 90.

Zi, Y. (2010). Objective measurement in the psychological science field—Traits and developmental trends of Rasch model. *Psychological Science Progress,* 1302.

Chapter 12
Development and Psychometric Evaluation of an Integrative Thinking Mode Scale Utilizing the Rasch Measurement Model

Zhuanmao Song and Bo Jiang

Abstract *Objectives* There is need for a validated simple instrument to quantify the mode of integrative thinking to aid integrative practice of being, and the purpose of the present study was to develop a scale (questionnaire) to measure the mode of integrative thinking in individuals utilizing the Rasch measurement model as a tool of psychometric analysis. *Method* Based on an literature review and feedback from five experts in psychology, a questionnaire consisting of 53 items and containing 8 factors was first created and then a survey completed by 538 college students (gender: 296 males, and 242 females; grade: 217 freshmen, 121 sophomores, 116 juniors, and 84 seniors; age: 18–26 years with the mean age being 20.15 years, standard deviation = 0.6) was submitted to Rasch model analysis. *Results* A final scale with 48 items containing 8 factors (including a factor about response effectiveness with 4 items) was provided that is internally consistent and reliable measures of the mode of integrative thinking for participants after amendments. The items of the final scale have good fit for Rasch model, and the scale has high PSI providing statistical evidence of reliability. The scale could benefit from the items dealing with high levels of the mode of integrative thinking. *Conclusion* The scale appears to be a valuable tool for the assessment of the mode of integrative thinking and may be an attractive option for researchers, clinicians, and managers seeking to measure the levels of the mode of integrative thinking in individuals. In addition to its valid theoretical structure and sound psychometric properties, the scale has advantages over other ways to evaluate the mode of integrative thinking of being, as it is conducted objectively providing evidence of content validity.

Keywords Mode of integrative thinking · Rasch model · Scale · Development

Z. Song (✉)
School of Education, Huazhong University of Science
and Technology, Wuhan 430074, China
e-mail: profsong@126.com

B. Jiang
Guangzhou University, Guangzhou 510006, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_12

12.1 Introduction

This study was part of a larger project funded by Bureau of Education of Guangzhou Municipality, examining college students' differences in creativity based on a view of integrative thinking to improve training of bright people in higher education (Song and Wang 2011–2014), while the data reported were about developing an objective tool that can effectively measure persons' differences in mode of integrative thinking utilizing the Rasch model as a tool of psychometric analysis.

12.1.1 The Whys to Develop Integrative Thinking Scale and Utilize Rasch Model

As an important aspect of cognitive mode, modes of thinking are referred to the preference of "mind" to use wisdom, capability, knowledge, and skills to solve problems successfully (Tullett 1996). Thus, the mode of integrative thinking can be conceptualized as a preference of "mind" to interact with self, world, and others in predicaments that are unstructured, ill-structured, ambiguous, or calling for the integration of disparate knowledge structures and behavioral patterns into a single cognitive schema and behavioral action plan (Moldoveanu 2005).

The integrative mode of being is characterized by effectively solving complex problems, integrating different things or concepts together to create a new thing or idea of which the functions are better than that of the original. And it helps people get successful resolutions of the tensions between the need to learn and adapt and the need to act decisively and purposefully through interacting with self, world, and others (Moldoveanu 2005). In fact, integrative thinking is not only the important resources bolstering integrative activities of being (Westra and Rodgers 1991; Rosch 1998; Lima et al. 2004) but also the essence of creation (Sill 1996).

However, the mode of integrative thinking is neither purely cognitive function nor a function of personality but a trait in between them (Sternberg and Grigorenko 1997). Its level is an individual-difference variable in human integrative performance and has long occupied the minds of many scholars. Since 1980s, there has been a proliferation of literature in the area of theories and models of integrative thinking that has become stagnant partially because of a lack of objective scale that can effectively measure persons' differences in the mode of integrative thinking and then be readily accepted by others. But the mode of integrative thinking is increasingly valued along with growing of the need for solving complex problems and integrative activities in social economy, politics, culture, and scientific research. Therefore, there is need for a validated simple instrument to quantify the mode of integrative thinking in individuals to aid integrative practice of being. It is very important to develop an objective scale that can effectively measure persons' differences in the mode of integrative thinking, with which people might evaluate and diagnose the mode of integrative thinking and take educational intervention measures to enhance the efficiency of integrative practice of being.

Meanwhile, objective psychological assessment scales are mainly developed through statistical methods such as principal component analysis, cluster analysis that are based on statistical average, and standard deviation, which need to assume that the measurement tools are isometry. However, measurement tools commonly used to measure people's ability, attitude, and personality do not meet the condition (Stevens; Embretson 1996), and the way to accumulate raw scores (including the linear transformation, such as T scores) is invalid.

What's worse, in the traditional procedure of developing psychological assessment scales, subjects' abilities or levels of traits are defined by the test scores, while item difficulty is defined by the right rate. If the test is very simple, the subject's score is high, and the ability is strong. Conversely, if the test is difficult, the subject's score is low, and the ability is weak. The subject' s level of ability depends on the nature of the test, and this is the so-called test-dependent. By the same token, if the subject's ability is weak, right rate is low, so the item difficulty is high. On the other hand, if the subject's ability is strong, right rate is high, the item becomes very simple. Item difficulty depends on the characteristics of the subject sample, and this is the sample-dependent. In short, both to estimate subject's ability and to estimate item difficulty interfere each other, and of course, there is no "objective" or equal measure. When the problems mentioned earlier can't be effectively solved, the test scores refer to little information and then the subsequent analysis is also problematic.

Rasch model (Rasch 1960) is helpful to get rid of the predicament. This model is designed to establish a set of objective standards that fit the measurement in the field of social science with the benchmark in the field of natural science. Rasch Model is a revolutionary breakthrough to the traditional method of measurement scale development (Bond and Fox 2007). It helps the psychological and educational measure to get out of the nonobjective isometric dilemma and achieve the equidistant gauge in objective measurement being different from IRT (Yan 2010). For Rasch model, both the individual ability and the item difficulty are put in the same scale with data being fit for the model as the premise, namely, a characteristic of the Rasch model is that items are thought to vary just in terms of their item difficulty, by which the test-dependent and the sample-dependent mentioned earlier are thus overcome. And the objective and reliable measurements are ensured with the corresponding improvement in the information analysis of test items. That's the whys to utilize Rasch model to develop integrative thinking scale.

12.1.2 The Theoretical Construction of Integrative Thinking Mode Scale

What are the features of the mode of integrative thinking? A theoretical way could be chosen to solve this problem through which the factorial structure of integrative thinking mode scale could be set up.

From the view of epistemology, the modes of thinking are considered by stance (attitude), tools, and skills of thinking and knowledge of meta-cognition (Tian 1990). Empirically, the stances of integrative thinkers are characterized by pursuing integration and functional optimization of things (Rosch 1998; Martin 2009), believing that the whole is greater than the sum of its parts (Rosch 1998) and dialectically looking upon things and their relationship (Labouvie-Vief and Diehl 2000). Integrative thinkers hold the opinion that all the things in the world interact with each other or become potential resources and have complementary relations of function even if they seem to be disparate or separate (Yan and Arlin 1999). That is to say, the mode of integrative thinking should implicate the factors of pursuit of integrative optimization, function-orientation, and resource-orientation

In terms of technology and tools used in the process of thinking, an integrative thinker involves in analysis and synthesis (Sill 1996). Sill holds that synthesis can be concentrated on the meaning of an integrative thinking skill, although the word may be used in a different sense. Klein (1996) thinks that an integrative thinker is used to utilizing bisociative thinking during solving a complicated problem. In the light of Koestler's view, bisociative thinking is a different conception from associative thinking. Associative thinking works within the confine of a single matrix, and while associative thinking may be very complex, bisociative thinking works at the intersection of distinctly separate matrices, whether those matrices are simple or complex. Koestler (1964) also defines integrative thinking as a multidimensional affair as is confirmed by Dixon, Martin, Muraven and Baumeister. Dixon (2006) thinks an integrative thinker is used to systematize according to his investigation results. Meanwhile, Martin (2009) thinks that an integrative thinker is used to utilizing statistical induction based on his investigations as well as reconcile that has been proved by Klein (1996) and meta-cognitive control that Muraven and Baumeister (2000) have authenticated. Through the use of these technologies and means, an integrative thinker can find the relative things and their relations, with making them stand out. And from the overall perspective of functional optimization, he or she has to do with making something complete by the addition of components that are lacking, and to accomplish this it may be necessary for him or her to draw on elements that are in seemingly foreign realms (Rosch 1998). Therefore, an integrative thinker would tend to utilizing analysis (including comparison, evaluation), bisociation, statistical induction, systematization, and cognitive regulation. They think in a brand new way to explore relations among things and to solve problems. Thus, the factors of integrative thinking scale become visible.

Based on the literature review and information about the features of integrative thinking mentioned earlier, a theoretical frame of the scale structure about the mode of integrative thinking is proposed (see Diagram 12.1).

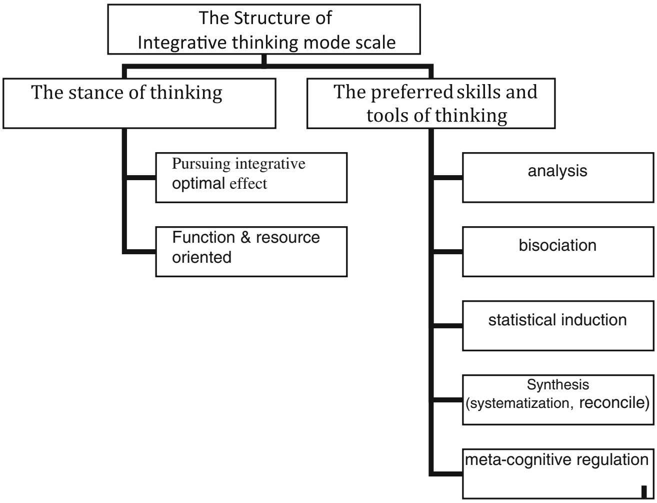

The Structure of
Integrative thinking mode scale
The stance of thinking
The preferred skills and tools of thinking
Pursuing integrative optimal effect
Function & resource oriented
analysis
bisociation
statistical induction
Synthesis (systematization, reconcile)
meta-cognitive regulation

Diagram 12.1 The structure of integrative thinking mode scale

12.2 Methods

12.2.1 Participants

The sample comprised 538 Chinese college students (296 boys and 242 girls) from five universities in Guangzhou, South of China, for the pilot study. In addition, 217 of the participants were freshman, 121 were sophomore, 116 were junior, and 84 of the participants were senior. The participants ranged from 18 to 23 years in age, and the mean age of the sample was 20.15 years (standard deviation [*SD*] = 0.6). Finally, it should be noted that a random sampling procedure was not used in this study; however, the participants seemed to be proportionally representative of the broader university population with respect to gender.

12.2.2 Instruments

On the basis of the literature review, 60 items were first compiled for the Integrative Thinking Mode Scale (ITMS) that was evaluated by five senior researchers from the Psychology department of a large university in Guangzhou, south of China. The five senior researchers are all senior faculty or expert psychologists in the area of psychological measurement who were paid for evaluating the fit of each item, of which 7 items were proposed to be dropped off.

Students complete the 53-item ITMS (see Appendix for a list of item as translated to English) within a class hour. All the items are self-report designed to measure the preference of integrative thinking on a 5-point Likert-type scale. That is, each student is required to provide information about their modes of integrative thinking over the past several months. Items on the ITMS are rated using a 5-point Likert-type scale of 1 = "**Definitely No**"; 2 = "**Possibly No**"; 3 = "**Maybe**"; 4 = "**Possibly Yes**"; and 5 = "**Definitely Yes**". Scale scores are computed by creating a sum score of item ratings, which is transformed to a *T* score (*M* = 50, *SD* = 10) in order to classify children at risk of developing behavioral and emotional problems. To be considered "at risk," a child must have a BESS *T* score equal to or greater than 60.

5 = Definitely Yes; 4 = Possibly Yes; 3 = Maybe; 2 = Possibly No; 1 = Definitely No

The ITMS is originally composed of 8 aspects:

① Pursuing integrative optimal effect: 8 items, e.g.,
Things opposing each other can be unified together.

② Function- and resource-oriented: 9 items, e.g.,
The critical to succeed in work is to mobilize all aspects of resources,

③ Bisociation: 8 items, e.g.,
Ideas as "I am not rich, but very kind" often emerge in my mind.

④ Analysis: 5 items, e.g.,
Handling difficult problems, the first thing is to find out the crux of them.

⑤ Statistical induction: 4 items, e.g.,
I often find out implicit rules through the analysis of observed count of events.

⑥ Synthesis (systematization, reconcile): 9
Facing complex situation, I would be easy to sort out my own ideas.

⑦ Meta-cognitive regulation: 6 items, e.g.,
In solving a problem, I realize what method I am utilizing.

⑧ Response effectiveness: 4 items, e.g.,
I have never told a lie.

12.2.3 Tools and Logic Basis for Data Analysis

WINSTEPS software (v.3.81.0; Linacre 2014) was used for all analyses in the current study, through which the Rasch Rating Scale Model (RSM) information is provided about the difficulty and ability levels estimated from the model and the necessary assumptions as well as the information about model-data fit is provided.

The logic basis for data analysis was the assumption of Rasch models as follows: (1) construct unidimensionality, (2) a monotonic scale, and (3) the items that fit the Rasch model, as shown by fit indices within acceptable boundaries (Bond and Fox 2007). In addition, the RSM requires that each response category have a minimum frequency of 10, that the rating scale categories increase in difficulty of endorsement (named step values), and that the thresholds for each item are ordered.

12.3 Results

12.3.1 Unidimensional Test

A standard Principal Component Analysis offered by WINSTEPS was performed based on the scored observations, with the results showing that there are multidimensionality problems. According to Rasch model simulations, it is unlikely that the first contrast in the "unexplained variance" (residual variance) will have a size larger than 2. Here it is 3.2 although the variance explained by the first contrast is 13.5 %, less than the variance explained by the item difficulties 21.9 %. A secondary dimension in the data appears to explain more variance than is explained by the Rasch item difficulties. Below is the scale plot showing the relative sizes of the variance components (see Table 12.1). Thus, multidimensionality (subscales) analysis should be conducted, with the results being shown in Table 12.2.

Principal Components analysis offered by WINSTEPS (v.3.81.0) has been conducted one by one for all the seven subscales. The results are listed in Table 12.2.

According to the data listed in Table 12.2, the eigenvalue of the first contrast of each subscale is less than 2, and the raw variance explained by measures is 38–70.1 %. Based on the set of information, the ITMS may be said to have met unidimensionality (Raiche 2005). So the following analysis could be performed with Rasch model.

12.3.2 Measurement Reliability of Items and Persons

As traditional psychometric theory, the primary objective was to create a questionnaire made up of the smallest number of items that formed an instrument with reliable measurement properties in developing the analysis plan. According to Rasch model, measurement reliability of Items and persons is embodied by the

Table 12.1 Table of standardized residual variance (in eigenvalue units)

	Observed (%)		Expected (%)	
Total raw variance in observations	33.2	100.0		100.0
Raw variance explained by measures	9.2	27.7		28.9
Raw variance explained by persons	1.9	5.8		6.1
Raw variance explained by items	7.3	21.9		22.8
Raw unexplained variance (total)	24.0	72.3	100.0	71.1
Unexplained variance in 1st contrast	3.2	9.7	13.5	
Unexplained variance in 2nd contrast	2.3	6.9	9.5	
Unexplained variance in 3rd contrast	1.9	5.6	7.8	
Unexplained variance in 4th contrast	1.7	5.0	6.9	
Unexplained variance in 5th contrast	1.6	4.9	6.8	

Table 12.2 Principal-components analysis of residuals

The ab. name of subscale	Meaning subscale	Eigenvalue of the first contrast	Raw variance explained by measures
PIOE	Pursuing integrative optimal effect	1.7	46.7
FRO	Function-oriented and resource-oriented	1.9	38.0 %
ANA	Analysis including comparison and evaluation	1.9	48.3 %
BIS	Bisociation (by Koestler), to create novel meanings	1.8	38.0 %
SIN	Statistical induction, to find meanings from increase or decrease in quantity	1.7	70.1 %
SYN	Synthesis including systematization, reconcile	1.7	46.2 %
MCR	Meta-cognitive regulation	1.4	41.6 %

separation of items and persons. The data listed in Table 12.3 show that all the separation indices of items are very high, with the value being 5.97–8.27 and more than 6 except for that of the subscale SIN (Wright and Masters 1982), that the measurement reliability of items is 0.89–0.97, while the separation indices of persons are 1.68–2.38, and the measurement reliability of persons is 0.71–0.82 and that the Cronbach Alpha is 0.79–0.90. In terms of the information mentioned earlier, the internal consistency of the seven subscales is good.

12.3.3 A Rasch Test of the ITMS Category

Next, the practicality of the 5-point rating scale was examined. The results of a Rasch test for the ITSM category are listed in Table 12.4. The data in the second column show that each category has enough cases (>10) to meet the count of Rasch model estimate, and the measure mean of each category is monotonic scale. However, the mode of five response categories doesn't meet other standards of Rasch model as the threshold of the category is so small that many of them are less

Table 12.3 Measurement reliability of Items and persons

The ab. name of subscale	Item count	Item separation	Item reliability	Person separation	Person reliability	Cronbach alpha
PIOE	8	8.03	0.91	2.16	0.82	0.90
FRO	10	7.46	0.97	1.93	0.79	0.83
ANA	5	8.27	0.93	2.03	0.80	0.87
BIS	6	6.18	0.90	2.38	0.84	0.80
SIN	4	5.97	0.89	2.09	0.80	0.79
SYN	9	7.20	0.95	1.68	0.78	0.81
MCR	7	3.67	0.91	2.09	0.71	0.85

Table 12.4 Statistical test results of category

	Observed count	Category measures	Infit	Outfit	Structure measures	Coherence (%)	
			MNSQ	MNSQ		M → C	C → M
PIOE							
Definitely no	437	−2.31	1.53	1.28	None	40	18
Possibly no	979	−0.14	0.81	0.79	−1.35	42	46
Maybe	1076	−0.07	1.18	0.97	−0.26	49	39
Possibly yes	1091	1.09	0.89	0.85	0.1	44	67
Definitely yes	722	2.47	0.93	0.98	1.51	64	15
FRO							
Definitely no	547	−2.73	1.13	1.15	None	44	25
Possibly no	1223	−1.07	0.98	0.99	−1.45	41	31
Maybe	1345	0.08	1.20	1.39	−0.24	50	38
Possibly YES	1363	1.08	0.67	0.63	0.28	42	69
Definitely yes	903	2.70	0.92	0.94	1.40	71	16
ANA							
Definitely no	279	−2.83	1.22	1.28	None	43	55
Possibly no	629	−1.29	0.71	0.68	−1.55	51	35
Maybe	641	−0.01	0.87	0.81	−0.44	57	44
Possibly yes	685	1.35	0.88	0.85	0.39	46	52
Definitely yes	458	2.87	1.26	1.22	1.60	71	20
BIS							
Definitely no	278	−3.24	1.29	1.22	None	41	12
Possibly no	652	−1.32	0.56	0.57	−2.01	53	68
Maybe	803	0.13	1.08	1.19	−0.40	40	38
Possibly yes	772	1.36	0.81	0.74	0.85	43	67
Definitely yes	731	2.93	1.12	1.12	1.56	66	33
SIN							
Definitely no							
possibly no	219	−2.83	1.22	1.28	None	43	55
Maybe	489	−1.29	0.71	0.68	−1.55	51	35
Possibly yes	538	−0.01	0.87	0.81	−0.44	57	44
Definitely yes	545	1.35	0.88	0.85	0.39	46	52
	361	2.87	1.26	1.22	1.60	71	20
SYN							
Definitely no	493	−2.68	1.34	1.37	None	39	26
Possibly no	1101	−1.02	0.69	0.59	−1.68	43	37
Maybe	1211	−0.32	0.85	0.82	−0.38	47	38
Possibly yes	1226	1.07	0.91	0.88	0.43	49	47
Definitely yes	809	2.69	1.19	1.23	1.59	67	19
MCR							
definitely no	383	−2.79	1.27	1.23	None	41	34
Possibly no	856	−1.19	0.73	0.59	−1.58	57	29
Maybe	941	−0.08	0.82	0.77	−0.41	49	41
Possibly yes	953	1.31	0.85	0.83	0.37	44	42
Definitely yes	632	2.07	1.24	1.26	1.55	66	22

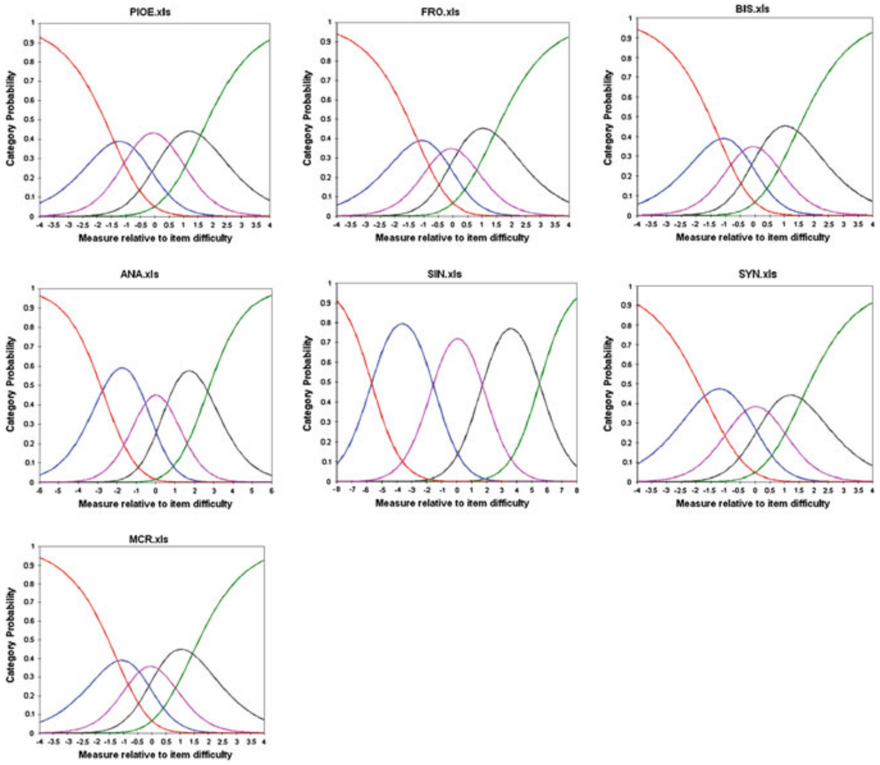


Fig. 12.1 Probability category curves

than 0.8. That is to say, there are little differences estimated between the persons to choose “possible yes” and the ones to choose “maybe”, which suggest that the participants cannot effectively distinguish the response categories. Therefore, the response categories of the original scale need to be restructured in different ways.

Having tried a variety of other categories, there is a mode of response categories found that can conform to the standard that incorporating categories 2, 3, and 4 (responding at 3, 13,335). Through the reconstruction of the response categories, the unidimensionality of the scale and the threshold of the category are improved, with all categories showing proper fit as shown in Fig. 12.1.

Figure 12.1 shows the probability category curves of seven subscales. From the information shown in Fig. 12.1, each of the category probabilities of seven subscales is more than zero and changes as the estimated measures do.

Table 12.5 The results of the DIF test in gender

DIF contrast 反差		
Subscale	Min	Max
PIOE	−0.34	0.51[#]
FRO	−0.35	0.49
ANA	−0.31	0.48
BIS	−0.29	0.34
SIN	−0.33	0.37
SYN	−0.26	0.70[#]
MCR	−0.40	0.36

[#] It means that the value is not fit for the standards of Rasch model and some items need to be adjusted

Table 12.6 The results of the DIF test in gender

DIF contrast 反差		
Subscale	Min	Max
PIOE	−0.31	0.49
FRO	−0.35	0.49
ANA	−0.31	0.48
BIS	−0.29	0.34
SIN	−0.33	0.37
SYN	−0.28	0.47
MCR	−0.40	0.36

12.3.4 The Differential Item Functioning Test in Gender

Differential item functioning (DIF) is regarded as a kind of validity to test the quality of a scale, as the quality of a scale is often called in question when the items show outstanding properties in DIF. To test the DIF of the ITMS, the measure in gender dimension of participants is taken into account here.

The results are shown in Table 12.5. As can be seen at a first glance, two subscales have been flagged for DIF in terms of the standards of Rasch model. To look into the reasons, a test for each of the items in subscales SYN and PIOE was conducted by Rasch model, which a few items including Items 35 (My attention is easily affected by the surrounding atmosphere), 37 (My idea about a controversial topic is easily influenced by the person recently talked with), 42 (I think that it's against the success of a task when decomposing a whole goal), and 53 (I like to ask myself whether to achieve desired goals or not while doing anything) show more than negligible DIF effects. Having deleted the four items, the results in Table 12.6 show that all of the subscales are fit for the standards of Rasch model and that the DIF is less than 0.5.

12.3.5 Item Statistics

The results of item statistics are listed in Table 12.7, of which the data in Columns 4 and 5 show that the items except Item 46 are fit for Rasch model (Linacre 2014), with the infit mean squares (MNSQs) being between 0.5 and 1.5. Furthermore, the data in Column 6 show that the values of point measure-A correlation are more than 0.4, which means that there are rational relations among the items except Item 46. In the light of information mentioned earlier, Item 46 is suggested dropping off.

12.4 Discussion

The present study aims to develop a practical and objective tool to estimate the mode of integrative thinking in individuals and to aid integrative practice of being. Based on a literature review and the information about the features of integrative thinking, a theoretical frame about the mode of integrative thinking was proposed, with seven domains defined. The seven domains implied that the mode of integrative thinking consisted of seven main factors, i.e., ① pursuing integrative optimal effect, ② function- and resource-oriented, ③ bisociation, ④ analysis, ⑤ statistical induction, ⑥ synthesis (systematization and reconcile), and ⑦ meta-cognitive regulation. For each factor, 6–10 items were designed to survey the degree of mode of integrative thinking.

The opinions of expert psychologists in the area of psychological measurement were accepted in designing first items, and the item reduction process followed a rigorous methodology by the Rasch measurement model, together with careful monitoring of content. Since Rasch methodology provides an objective test of the quality of each item's fit to a unidimensional model that contains all the items, the composition of the final item set was driven more by the student's responses. Items with poor measurement properties were dropped off while preserving broad coverage of the different effects of the mode of integrative thinking. Of course, it is necessary to balance its weaknesses and strengths against its overall contribution while deciding whether to include or exclude an item during questionnaire development. Fit was examined for each item on the questionnaire along with Rasch reliability indices for both persons and items. For each item, the frequency of responses to every category was examined to ensure that there were sufficient numbers of participants per category. Also, scale steps were examined to verify that as higher levels on the rating scale were selected, there was an increase corresponding to the amount of scores observed. Finally, item statistics was conducted (Bond and Fox 2007; Linacre 2014).

The final content covers all of the seven factors mentioned earlier and a factor about response effectiveness. Items considered to survey the same factor were organized in a defined subscale so that a scale (questionnaire) that consisted of

Table 12.7 The statistical test results of items

Scales and items	Model measures	Model S.E.	Infit MNSQ	Outfit MNSQ	Point measure-A CORR.
PIOE					
3. Things opposing each other can be unified together	0.49	0.08	1.44	1.43	0.48
50. I like to associate a thing with another entirely different one	0.44	0.08	0.71	0.71	0.55
5. A good method is integrated by absorbing the advantages of other ones	0.29	0.08	0.91	0.94	0.47
51. I get often excited at finding the meaning between different things	0.23	0.1	0.79	0.77	0.64
52. I'm used to thinking of all the factors to seek the best solution to a problem	−0.10	0.11	1.10	1.05	0.48
4. I don't spend time to think of new ways if there being reluctant one available	−0.44	0.10	0.73	0.74	0.44
2. I don't like making things detailed, so as not to make things complicated	−0.90	0.10	1.41	1.31	0.41
FRO					
12. The critical to succeed in work is to mobilize all aspects of resources	0.84	0.10	1.14	1.22	0.50
14. Many things seemingly unrelated imply usefulness actually	0.77	0.10	1.40	1.47	0.72
9. The most important is the abilities as roles while looking for team members	0.44	0.12	0.82	0.79	0.57
7. I like thinking of the function of things	0.31	0.08	1.0	1.0	0.55
8. It depends on the usefulness for things to be got together	0.11	0.08	0.81	0.81	0.71
6. The key to choose partners is to see if they can give a help	−0.18	0.08	0.78	0.78	0.53
10. I believe that everything is potentially useful	−0.34	0.06	0.83	0.90	0.53
13. I like caring about the information related to the things to do	−0.72	0.05	0.80	0.79	0.48
11. I like taking people or things contacted with as useful resources	−1.12	0.08	1.25	1.25	0.41
ANA					
24. Handling difficult problems, the first thing is to find out the crux of them	0.55	0.06	1.27	1.24	0.63
25. Finding others superficially debating a point of view, I would be worried	0.54	0.06	1.07	1.03	0.74

(continued)

Table 12.7 (continued)

Scales and items	Model measures	Model S.E.	Infit MNSQ	Outfit MNSQ	Point measure-A CORR.
26. Presenting a view, I'm used trying to find out what evidence to support it	0.16	0.11	0.85	0.9	0.58
27. Hearing others stating opinions, I'm used to trying to find out the point of view	−0.06	0.11	1.0	0.98	0.57
28. Refuting opinions of others, you must give convincing reasons	−1.03	0.11	0.87	0.87	0.62
BIS					
22. Ideas as "I am not rich, but very kind" often emerge in my mind	0.49	0.06	1.13	1.14	0.47
21. Finding a person's advantages, I can't help to think of their disadvantages	0.41	0.06	1.02	1.0	0.62
23. I often think of the beneficial aspects of failed events	0.16	0.10	1.31	1.35	0.51
19. I often talk about a thing from the opposites	0.14	0.10	0.83	0.83	0.66
18. Meeting a new thing, I like thinking of what it's like to be on the back	−0.20	0.10	0.87	0.95	0.54
20. I'm used to associating with my advantages while analyzing my weaknesses	−0.42	0.10	1.0	0.82	0.55
16. I'm used to thinking of the opposites of a thing	−0.58	0.10	0.88	0.73	0.60
17. I like thinking about solutions to a problem from the opposite	0.61	0.08	0.92	0.88	0.56
SIN					
32. For the comparison of things, I often try to use quantitative methods	1.47	0.05	0.73	0.72	0.85
30. I like using statistics methods to find out the relationship between things	1.29	0.05	0.64	0.64	0.86
31. I like processing objects quantitatively in order to enhance the objectivity of evaluation	−0.38	0.10	1.09	1.10	0.77
29. I often find out implicit rules through the analysis of observed count of events	−3.14	0.10	1.43	1.48	0.43
SYN					
36. I like firstly analyzing the focus of questions, then answering them	1.12	0.07	1.35	1.30	0.53
34. Facing complex situation, I would be easy to sort out my own ideas	0.82	0.07	1.37	1.46	0.64

(continued)

Table 12.7 (continued)

Scales and items	Model measures	Model S.E.	Infit MNSQ	Outfit MNSQ	Point measure-A CORR.
40. I'm good at designing the process of study or work according to a target	0.15	0.07	1.10	1.13	0.70
38. I often find the underlying laws or useful information in messy materials	−0.19	0.10	0.81	0.85	0.70
39. I often summarize an opinion for everyone to accept from the dispute ones	−0.58	08	0.77	0.79	63
41. I'm easy to take the opinions of others into my thoughts	−0.99	08	0.66	0.70	51
MCR					
49. I would ask myself whether there are better ways of doing things after a finish	0.43	0.08	1.04	1.12	0.54
45. I would try to change my way of doing things in order to adapt to different conditions and task requirements	0.27	0.08	1.13	1.20	0.42
44. In solving a problem, I realize what method I am utilizing	−0.12	0.11	1.22	1.25	0.61
47. Before I do something, I'm used to setting some clear goals	−0.23	0.11	0.91	0.97	0.47
48. I'm used to conceiving a variety of methods that I know when, where and why use them to do things	−0.48	0.11	0.84	0.90	0.48
46. I would change the original method and strategy in the process of doing things, if necessary	−0.55	0.11	1.68	1.73	0.38

seven subscales was developed with the amount of items being 48, together with four items for response effectiveness.

According to the examined results, the quality of fit to a Rasch unidimensional model suggests that all the subscales have true interval scaling properties. Based on data from four grades of five universities, tests of internal consistency show that it provides a reliable measure of overall mode of integrative thinking severity from the students' experience in universities. This should ensure that it is applicable for worldwide use.

There are two limitations of the present study. One is that the reliability and validation findings are based on data from the college students, owing to the requirements of the project funded by Bureau of Education of Guangzhou Municipality. The other is lack of criterion validity. Further study is required.

In summary, the scale appears to be a valuable tool for the assessment of the mode of integrative thinking and may be an attractive option for researchers, clinicians, and managers seeking to measure the levels of the mode of integrative thinking in individuals. In addition to its valid theoretical structure and sound psychometric properties, the scale has advantages over other ways to evaluate the mode of integrative thinking of being, as it is conducted objectively providing evidence of content validity.

Appendix: Items for the ITMS

1. Sometimes I put off what to be done until next day.
2. I don't like making things detailed, so as not to make things complicated.
3. Things opposing each other can be unified together.
4. I don't spend time to think of new ways if there being reluctant one available.
5. A good method is integrated by absorbing the advantages of other ones.
6. The key to choose partners is to see if they can give a help.
7. I like thinking of the function of things.
8. It depends on the usefulness for things to be got together.
9. The most important is the abilities as roles while looking for team members.
10. I believe that everything is potentially useful.
11. I like taking people or things contacted with as useful resources.
12. The critical to succeed in work is to mobilize all aspects of resources,
13. I like caring about the information related to the things to do.
14. Many things seemingly unrelated imply usefulness actually.
15. I occasionally think of some bad things.
16. I'm used to thinking of the opposites of a thing.
17. I like thinking about solutions to a problem from the opposite.
18. Meeting a new thing, I like thinking of what it's like to be on the back.
19. I often talk about a thing from the opposites.
20. I'm used to associating with my advantages while analyzing my weaknesses.
21. Finding a person's advantages, I can't help to think of their disadvantages.
22. Ideas as "I am not rich, but very kind" often emerge in my mind.
23. I often think of the beneficial aspects of failed events.
24. Handling difficult problems, the first thing is to find out the crux of them.
25. Finding others superficially debating a point of view, I would be worried.
26. Presenting a view, I'm used trying to find out what evidence to support it.
27. Hearing others stating opinions, I'm used to trying to find out the point of view.
28. Refuting opinions of others, you must give convincing reasons.
29. I often find out implicit rules through the analysis of observed count of events.
30. I like using statistics methods to find out the relationship between things.

31. I like processing objects quantitatively in order to enhance the objectivity of evaluation.
32. For the comparison of things, I often try to use quantitative methods.
33. Sometimes I really want to swear at people.
34. Facing complex situation, I would be easy to sort out my own ideas.
35. My attention is easily affected by the surrounding atmosphere.
36. I like firstly analyzing the focus of questions, then answering them.
37. My idea about a controversial topic is easily influenced by the person recently talked with.
38. I often find the underlying laws or useful information in messy materials.
39. I often summarize an opinion for everyone to accept from the dispute ones.
40. I'm good at designing the process of study or work according to a target.
41. I'm easy to take the opinions of others into my thoughts.
42. I think that it's against the success of a task when decomposing a whole goal.
43. I have never told a lie.
44. In solving a problem, I realize what method I am utilizing.
45. I would try to change my way of doing things in order to adapt to different conditions and task requirements.
46. I would change the original method and strategy in the process of doing things, if necessary.
47. Before I do something, I'm used to setting some clear goals.
48. I'm used to conceiving a variety of methods that I know when, where and why use them to do things.
49. I would ask myself whether there are better ways of doing things after a finish.
50. I like to associate a thing with another entirely different one.
51. I get often excited at finding the meaning between different things.
52. I'm used to thinking of all the factors to seek the best solution to a problem.
53. I like to ask myself whether to achieve desired goals or not while doing anything.

References

Bond, T. G., Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (pp. 101–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Dixon, R. (2006). Integrative thinking: Building personal working models of psychology that support problem-solving. The Higher Education Academy Psychology Network. http://www.Psychology.Heacademy.ac.

Embretson Equidistant Gauge, S. E. (1996). Item response and found models and spurious interaction effects in the factorial ANOVA designs. *Applied Psychological Measurement, 20*, 201–212 (Likert scale).

Klein, J. T. (1996). Crossing boundaries: Knowledge, disciplinarities, and interdisciplinarities. Charlottesville, VA: University of Virginia Press.

Koestler, A. (1964). The act of creation. New York: Macmillan.

Labouvie-Vief, G., & Diehl, M. (2000). Cognitive complexity and cognitive-affective integration: Related or separate domains of adult development? *Psychology and Aging, 15*, 490–504.

Lima, M., Koehler, M. J., Spiro, R. J. (2004). Collaborative interactivity and integrated thinking in brazilian business schools using cognitve flexibility hypertexts: The PANTEON PROJECT. *Journal of Educational Computing Research, 31*(4), 371–406.

Linacre, J. M. (2014). Facets—Rasch measurement computer program. Chicago, IL: MESA Press.

Martin, R. (2009). The opposable mind: Winning through integrative thinking [M]. Boston: Harvard Business Press.

Moldoveanu, M. (2005). Integrative thinking: The view from cognitive and social psychology. Circa 2005 AD: 1–15.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin, 126*(2), 247–259. (Washington, DC: American Psychological Association).

Paul, J., Rosch, M. D., FACP (1998). The integrative thinking: The essence of good medical education and practice. *The Integrative Physiological and Behavioral Science, 33*(2), 141–150.

Raiche, G. (2005). Critical eigenvalue sizes in standardized sesidual principal components analysis (PCA). *Rasch Measurement Transaction*, *19*(1), 1012.

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

Rosch, P. (1998). Integrative thinking: The essence of good medical education and practice. *Integrative Physiological and Behavioral Science, 33*(2), 141–140.

Sill, D. (1996). Integrative thinking, synthesis and creativity in interdisciplinary studies. *Journal of General Education, 45*(2), 129–151.

Sternberg, R. J., & Grigorenko, E. (1997). All cognitive styles, in style? *American Psychologists, 52*, 700–712.

Tian, Y. (1990). *Modes of thinking* (pp. 22–32). Fuzhou: Fujian Education Press.

Tullett, A. D. (1996). The thinking style of the managers of multiple projects: Implications for problem solving when managing change. *International of Project Management, 14*(5), 281–287.

Westra, B., & Rodgers, B. (1991). The concept of integration: A foundation for evaluating out-comes of nursing care. *Journal of Professional Nursing, 7*(5), 277–282.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis.Chicago: MESA Press.

Yan, Z. (2010). Objective measurement in the field of psychological science: Rasch model's point and development trend. *Progress in Psychological Science, 18*, 1298–1305.

Yan, B., & Arlin, P. (1999). Dialectical thinking: Implications for creative thinking. *Encyclopedia of Creativity* (Vol. 1, pp. 547–552). New York, NY: Academic Press.

Chapter 13
Assessment of the Psychometric Properties of the Tampa Scale of Kinesiophobia Adapted for People with Dystonia Using Rasch Analysis

I. Blackman, L. Bradnam and L. Graetz

Abstract Dystonia is a neurological movement disorder characterised by abnormal movements and postures and impacts on psychosocial health. One construct that is not well understood is movement-related fear or kinesiophobia. The Tampa Scale of Kinesiophobia (TSK) is used to assess kinesiophobia and activity avoidance in chronic low back pain and has been adapted for use in other musculoskeletal conditions. To date, there has been no adaptation of the TSK to a neurological movement disorder. *Aim* The aim was to develop a dystonia-specific version of the TSK (TSK-Dystonia) to assess kinesiophobia in people with dystonia. The objective was to investigate the psychometric properties of the TSK-Dystonia using Rasch analysis to determine whether it is able to assess a unidimensional construct of kinesiophobia. *Methods* A dystonia-specific TSK survey was constructed and refined by an expert group of experienced clinicians and people living with dystonia. The final version was offered online across Australia, New Zealand, United Kingdom, Europe and United States. Survey results were analysed by the partial credit version of the Rasch model. Specific attention was paid to exploring scale item fit, item threshold organisation and differentiated item function. *Results* One hundred and thirty two people (108 female) completed the survey. Rasch analysis identified four survey items as not demonstrating a unidimensional construct, with another two items having poor item threshold capacities. These six items were removed leaving an 11-item scale. The final TSK-Dystonia proved to be a reliable

I. Blackman (✉)
School of Nursing, Faculty of Medicine, School of Nursing & Midwifery, Nursing and Health Sciences, Flinders University, GPO Box 2100, 5001 Adelaide, SA, Australia
e-mail: Ian.blackman@flinders.edu.au

L. Bradnam
Discipline of Physiotherapy, University of Technology, Sydney, Australia

L. Graetz
School of Psychology, Faculty of Social and Behavioural Sciences, Flinders University, Adelaide, Australia

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_13

conjoint hierarchical scale, with category ordering, unidimensionality and interval consistency. However, participant gender and type of dystonia produced unwanted variance in two other scale items pointing to their possible removal, although this needs to be confirmed in a larger sample. *Conclusions* A Rasch analysis found an 11-item TSK-Dystonia was able to identify a unidimensional construct of fear avoidance in people with dystonia. In its present form, the TSK-Dystonia items could be reliably summed to provide a score determining the level of fear avoidance of an individual. The TSK-Dystonia could help to identify the presence of kinesiophobia in research and clinical dystonia populations and be used to measure improvements from targeted behavioural interventions.

Keywords Dystonia · Rasch · Tampa scale · TSK · Kinesiophobia · Activity-avoidance

13.1 Introduction

Misunderstanding and maladaptive beliefs related to pain have been a focus of research for over half a century. The development of a 'fear-avoidance' model to understand the psychosocial and behavioural reactions to injury and pain have improved recovery from injury, allowing people in chronic pain to regain their quality of life (Vlaeyen and Linton 2000). In part, maladaptive beliefs regarding their pain condition and the underlying cause leads to a learned response of activity avoidance known as fear avoidance or Kinesiophobia. The Tampa Scale of Kinesiophobia (TSK) (Kori et al. 1990) was developed to evaluate the construct of fear avoidance in people with chronic low back pain. The psychometric properties of the TSK assess the construct of fear avoidance by identifying maladaptive beliefs contributing to on-going disability (Vlaeyen and Linton 2000). The TSK is a 17-item questionnaire, with each item answered on a 4-point Likert-type scale ranging from 'strongly disagree' to 'strongly agree.' A total score is calculated after inversion of the individual scores of four 'reversed' items. The original TSK was modified for use in temporomandibular disorders (Visscher et al. 2010), fibromyalgia (Burwinkle et al. 2005), chronic fatigue syndrome (Nijs et al. 2004) and coronary heart disease (Back et al. 2013). Modifying the TSK to other disease states allows assessment of kinesiophobia arising from maladaptive beliefs that are appropriate to the particular condition.

Movement disorders are a set of neurological conditions, whereby the ability to move and participate in daily activity is impaired secondary to damage to the central or peripheral nervous system (Albanese 1990). Dystonia is a neurological disorder characterised by sustained or intermittent muscle contractions causing abnormal, often repetitive movements and postures (Albanese et al. 2013). There is a high degree of psychological stress in the form of depression and social withdrawal (Zetterberg et al. 2012; Zurowski et al. 2013). In people with dystonia, kinesiophobia could arise from the fear of triggering unsightly or painful dystonic postures by movement, leading to activity avoidance. In order to determine this, a dystonia-specific version of the TSK questionnaire is necessary. However, adaptations of original TSK to other conditions have encountered significant issues. First, the four reversed items only weakly correlate with the rest of the scale (Clark et al. 1996; Woby et al. 2005) and are often removed leaving a shortened 13-item version (Geisser et al. 2000). Second, the TSK does not always appear to provide a unidimensional construct of fear avoidance (Burwinkle et al. 2005; Woby et al. 2005). However, previous authors have used a range of methods to determine reliability of adapted scales, commonly principle components analysis and Cronbach's alpha (Roelofs et al. 2004; Burwinkle et al. 2005; Woby et al. 2005; Visscher et al. 2010; Bäck et al. 2012; Tkachuk and Harris 2012). These factor solutions are controversial and lack dimensional and construct consensus (Lundberg et al. 2009). A Rasch analysis can establish a scale that is unidimensional in construct to ensure scale reliability and therefore is superior to previous methods for this purpose (Leung et al. 2014). In support, Norwegian version of TSK for chronic back pain analysed using a Rasch analysis found the scale fitted a unidimensional construct of Kinesiophobia (Damsgard et al. 2007). Rasch analysis underpins a group of measurement models that identify latent traits, where scales for questions or survey (items) and the participants' subsequent scores are co-located along the same scale of the latent trait (Bond and Fox 2007). Scale item location (expressed as a range of difficulty or complexity) and participants' measures (expressed as participant ability) are analysed separately to produce estimates for each parameter, which are sample and item independent, respectively. Rasch modelling assesses the functioning of a scale in relation to response bias, dimensionality, response format, item content and appropriate targeting of the scale (Bond and Fox 2007; Pallant and Tennant 2007).

The aim of this study was to determine whether the TSK could be adapted to assess kinesiophobia in people with Dystonia. The objective was to assess the psychometric properties of an adapted version of the TSK (TSK-Dystonia) using Rasch analysis, exploring for survey item fit, person response reliability and differentiated item functioning (item bias) in order to establish a tool that identifies a unidimensional construct of Kinesiophobia in this population.

13.2 Methods

The original 17 items with 4 reversed-scored items of the TSK were used as a basis to develop the TSK-Dystonia (Appendix). The wording was adapted so that the word 'pain' was replaced with the word 'dystonia' and the questions modified to reflect beliefs surrounding this movement disorder, rather than a focus on pain. To ensure face validity of the proposed survey, a group of 'expert users' were consulted, comprising people with dystonia and experienced rehabilitation clinicians. Since pain can be a significant problem in many people with dystonia, the expert group was initially asked to complete both the proposed TSK-dystonia and the original TSK. They were then asked to judge whether the original version ‘captured’ any issues related to dystonia and whether there was a need for a dystonia-specific version. There was consensus that the TSK did not fully encapsulate the dystonia experience, and the TSK-dystonia scale was preferred. The wording of the proposed TSK-dystonia was refined in response to feedback and the amended version distributed to the expert group for a second time for final approval. The 17-item TSK-dystonia was distributed via dystonia special interest groups in Australia, New Zealand, United Kingdom, Europe and United States using an online survey platform. Demographics regarding location, gender, age, type of dystonia and years since onset were collected as part of the survey responses. Ethical approval was provided by the institutional ethics committee, and completion of the survey indicated consent by individuals.

13.3 Data Analysis

Results were first analysed using statistical analysis software (SPSS v20) to check for errors or omissions and ‘dystonia type’ recoded into focal and generalised dystonia. Data were then imported into Conquest software as developed by the Australian Council of Educational Research (Wu et al. 1997) and Winsteps 3.72 (Linacre 2011) to perform the Rasch analysis. The psychometric properties of the scale were determined using item fit statistics, individual item threshold order, differentiated item functioning and person fit statistics.

13.4 Results

13.4.1 Sample Demographics

There were 132 valid responses (108 female). The demographics of the population and time since diagnosis are provided in Table 13.1. The majority of respondents (107, 78 %) had been diagnosed with a form of focal dystonia (i.e. affecting an isolated part of the body such as the neck or hand). Twenty-three (17 %) indicated they had a form of generalised dystonia where more than one part of the body is affected.

13.5 Rasch Analysis

13.5.1 Item Fit Statistics

According to the Rasch model, all survey items are assumed to have equal discriminating power in identifying the underlying construct being estimated so that all

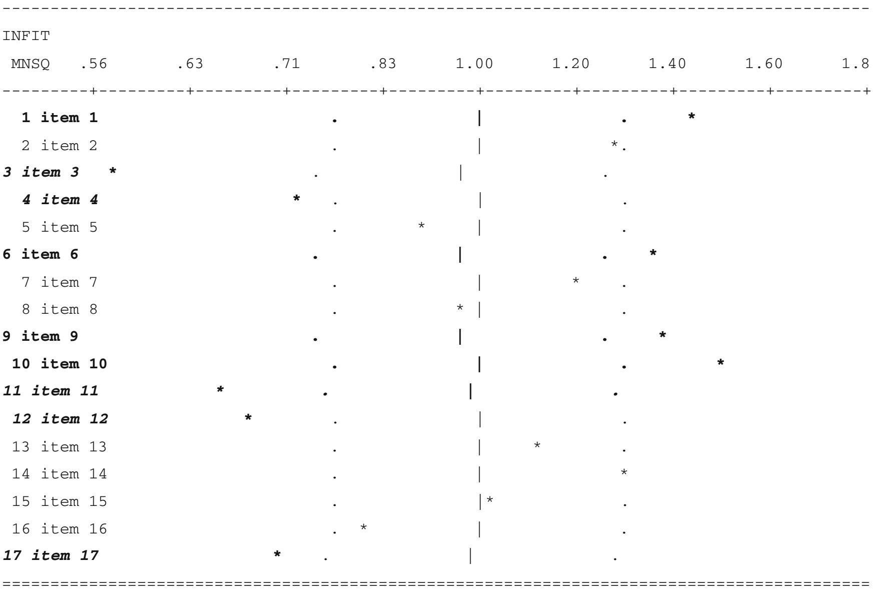
Table 13.1 Demographic data of the study participants

	n	(%)
Age		
18–30		
31–40		
41–50		
51–60		
61 or above		
Demographics		
Gender		
Male	108	21.0
Female	18	79.0
Place of residence		
Australia	51	37
United States of America	49	36
New Zealand	16	12
United Kingdom	9	7
Europe	7	5
Canada	3	2
Other	1	1
Dystonia-Related Characteristics		
Type of Dystonia		
Focal Dystonia[a]	107	78
Generalised Dystonia[b]	23	17
Length since Diagnosis		
Less than 2 years	20	14
2–10 years	67	49
11–20 years	31	23
Greater than 21	14	10
No answer	5	4

[a] Focal dystonia (i.e. affecting an isolated part of the body such as the neck or hand)

[b] Generalised dystonia where more than one part of the body is affected

survey items have unity or item fit statistics that are equivalent to a predetermined range, irrespective of which participants take the survey (Bond and Fox 2007). Figure 13.1 illustrates the degree to which the survey items used to determine participants' level of agreement about dystonia display unity or item fit in the current study. The vertical dotted lines in Fig. 13.1 represent the infit means square value ranges for the survey items and whether they are indeed unidimensional (all aligned to measure the same underlying variable, consensus), and then each survey items should fall between the (infit means square) values of greater than 1.30 and less than 0.75. Any items failing to meet these parameters are deemed not to



```
------------------------------------------------------------------------------------------
INFIT
 MNSQ   .56       .63       .71       .83      1.00      1.20      1.40      1.60      1.8
---------+---------+---------+---------+---------+---------+---------+---------+---------+
  1 item 1                        .              |              .      *
  2 item 2                        .              |             *.
3 item 3   *                    .              |              .
  4 item 4                    *   .              |              .
  5 item 5                        .        *     |              .
6 item 6                        .              |              .    *
  7 item 7                        .              |         *    .
  8 item 8                        .            * |              .
9 item 9                        .              |              .     *
 10 item 10                       .              |              .         *
11 item 11            *          .              |              .
 12 item 12              *        .              |              .
 13 item 13                       .              |     *        .
 14 item 14                       .              |              *
 15 item 15                       .              |*             .
 16 item 16                       .  *           |              .
17 item 17                  *    .              |              .
==========================================================================================
```

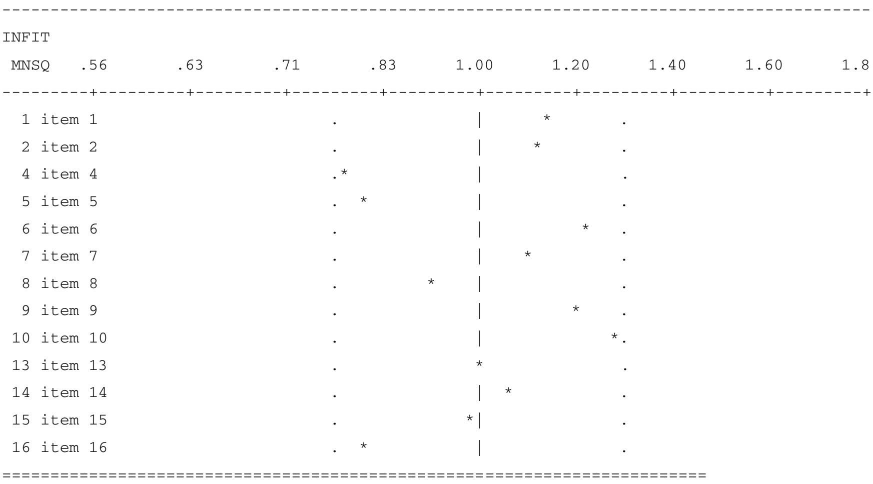
Fig. 13.1 Item fit statistics for initial TSK survey items. Note the survey items that have infit values of less than 0.75 confirm that responses to these items lack the variability of participant responses that the Rasch model predicts (items 3, 4, 11, 12 and 17). Conversely, items with outfit values above 1.30 suggest that participant responses were more haphazard or erratic than the Rasch model anticipated (items 1, 6, 9 and 10)

measure the same underlying construct as the remaining survey items and are removed from any further analysis. In Fig. 13.1, nine items do not meet this criterion for the predetermined range of the infit mean square values. Possible reasons for this could be that the wordings of items cause confusion for participants, and they are not answering that question in the way that is anticipated by the Rasch model (Blackman et al. 2006).

The survey items were re-estimated, progressively removing items deemed to be poorly fitting. Four items (3, 11, 12 and 17) were eventually removed. The final 13 items that were able to meet unidimensionality requirements are demonstrated in Fig. 13.2.

13.5.2 Individual Item Threshold Order

The TSK scale is a Likert-type scale that employs four consensus categories with three ordered thresholds. It is anticipated that each category of each item would be able to differentiate between the participants' capacity for consensus. It is assumed that participant responses to the categories in the survey would be different from



```
------------------------------------------------------------------------------------------
INFIT
 MNSQ   .56       .63       .71       .83      1.00      1.20      1.40      1.60      1.8
---------+---------+---------+---------+---------+---------+---------+---------+---------+
  1 item 1                        .              |      *       .
  2 item 2                        .              |     *        .
  4 item 4                        .*             |              .
  5 item 5                        .  *           |              .
  6 item 6                        .              |          *   .
  7 item 7                        .              |    *         .
  8 item 8                        .         *    |              .
  9 item 9                        .              |         *    .
 10 item 10                       .              |             *.
 13 item 13                       .              *              .
 14 item 14                       .              |  *           .
 15 item 15                       .             *|              .
 16 item 16                       .  *           |              .
=========================================================================
```

Fig. 13.2 Item fit statistics for revised TSK survey items, showing the final items that demonstrated unidimensionality

each other, and such differentiation of the categories (and their thresholds) would be ordered in all the tested items. Survey categories/thresholds that fail to show consensus differentiation from each other suggest low frequency counts of participant responses has occurred within each category (Wetzel and Carstensen 2014). However, questions with disordered item thresholds are not functioning adequately in this way (Andrich 2005). Figure 13.3 shows two items with disordered thresholds (Items 6 and 10).

Items 6 and 10 have threshold values that are equivalent to each other, meaning values for thresholds 1 and 2 are the same instead of sequentially increasing. This indicates that respondents are experiencing uncertainty in their consensus ability at the lower ends of the scale for these two questions. These two items are in effect only measuring the differences in participant consensus between two rather than three thresholds. These two items were removed from the final TSK scale and from further analysis on this basis.

13.5.3 Differentiated Item Functioning (DIF)

Rasch analysis determines whether invariance in the survey data violates unidimensionality of the construct being measured, in this case kinesiophobia in dystonia. Differentiated item functioning (DIF), also known as item bias, explores whether different members or subgroups within the cohort being measured differ markedly from the response patterns generated by the whole group (Bond and Fox 2007). If survey items differ significantly, then the items are biased and favour the

```
----------------------------------------------------------
    ITEM NAME       |SCORE MAXSCR|  THRESHOLD/S          |
                    |            |   1       2       3   |
----------------------------------------------------------
                    |            |                       |
6   item 6          |   338  411 |  -1.11   -1.11   -.34 |
                    |            |     .28     .28    .29|
                    |            |                       |
9   item 9          |   227  411 |   -.75    -.25    .87 |
                    |            |     .31     .29    .29|
                    |            |                       |
10  item 10         |   229  411 |   -.17    -.17   .18  |
                    |            |     .19     .19    .21|
----------------------------------------------------------
```

Fig. 13.3 TSK items showing disordered item thresholds for items 6 and 10

```
----------------------------------------------------------------------------------

                       Plot of Standardised Differences

   Easier to endorse for female                 Easier to endorse for male
   participants                                  participants

      -3         -2         -1          0          1          2          3          4
 -------+----------+----------+----------+-----------+----------+----------+----------+
 item 7           .                     |             *         .
 item 8           .                     |                       .          *
 item 9           .                  *  |                       .
==================================================================================
```

Fig. 13.4 Differentiated item functioning (gender) of TSK items showing item 8 is greater than 2 standard deviations from the mean, producing unwanted invariance

responses of one subgroup over other participants. In such cases, biased items should be removed to avoid unreliable study outcomes. The current study shows DIF based on the gender and dystonia type of groups of participants (three items in total). Figures 13.4 (gender, Item 8) and 13.5 (dystonia type, Items 1 and 3) shows item response patterns that are greater than two standard deviations (from the mean) based on the participants' attributes. Usually, under these circumstances, all items demonstrating DIF would be removed from further analysis. However, due to the

```
----------------------------------------------------------------------------------

                             Plot of Standardised Differences

  Easier to endorse for participants           Easier to endorse for participants
  With generalised dystonia                     with focal dystonia

      -3           -2           -1            0            1            2            3
 -------+------------+------------+------------+------------+------------+------------+
item 1            *.                         |                        .
 item 2             .               *         |                        .
 item 3             *                         |                        .
 item 5             .                        *|                        .
===========================================================================
```

Fig. 13.5 Differentiated item functioning (dystonia type) of TSK items showing items 1 and 3 were close to 2 standard deviations from the mean and could produce unwanted invariance

low distribution of males in the current cohort under study, Item 8 will be retained at present but should be further analysed in a group with better gender balance.

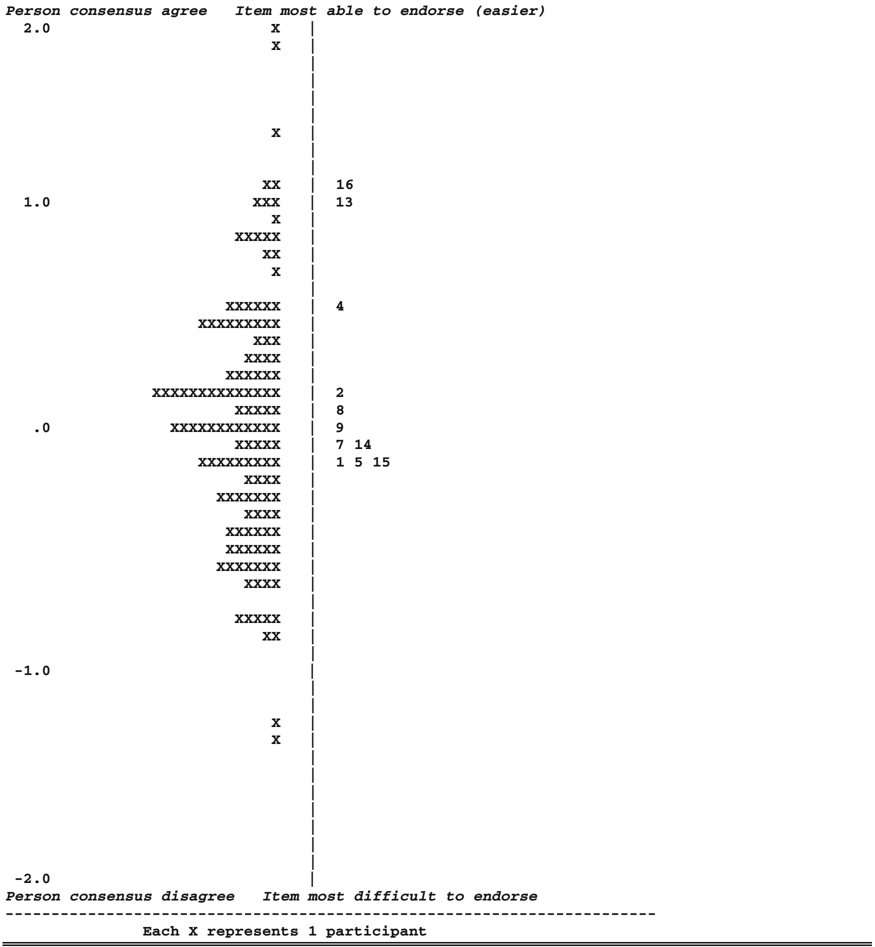
For dystonia type, Item 3 was already removed to generate the unidimensional construct. Item 1 is retained at present as there were fewer participants in the generalised dystonia group. Having explored individual item fit indices for threshold order and differentiated item function, it can be established that the remaining survey is unidimensional. The Item Separation Index was also most acceptable at 6.81, (reliability index of 0.98) which demonstrate the items tested for a wide range of severity in the TSK.

13.5.4 Person Fit Statistics

One remaining aspect to explore for reliability is the participants' response patterns to the survey items. Using a Person Separation Index (PSI), Rasch analysis has the capacity to identify whether participants tend to answer all the items by selecting the same option each time or tend to avoid using the categories located at either ends of the scale. The PSI analyses participant response patterns independent of the reliability of the survey items and should be above 2.0. For this survey, PSI = 2.60 (reliability of 0.84), indicating adequate replicability of the results. This result confirms that if this particular sample of participants were to complete the same survey again, a similar consensus placement on the Rasch scale can likely be achieved again (Yates 2005).

13.5.5 Item Consensus and Person Frequency Estimates Map

The Rasch model has the capacity to place all survey items on a hierarchical logit scale ranging from the easiest item for the participants to endorse to the most difficult. The Rasch model can do this conjointly, plotting against the participants' self-reported opinions that range from strongest agreement to strongest disagreement. Figure 13.6 depicts this hierarchy, with participants' consensus estimates extending vertically to the left of the dotted vertical line and the TSK item (known as item difficulty) to the right of it. In Rasch scale maps, the mean of the item consensus is set at zero (viewed as average ability to endorse), with the easiest items to agree with or endorse being located above the item mean and the harder survey items to endorse below the item mean. As the different survey items become easier to endorse, its level is demonstrated on the map relative to its positive logit value. In contrast, survey items rated as being more difficult to endorse are positioned on the



```
Person consensus agree   Item most able to endorse (easier)
  2.0                        X   |
                             X   |
                                 |
                                 |
                                 |
                                 |
                             X   |
                                 |
                                 |
                            XX   |  16
  1.0                      XXX   |  13
                             X   |
                         XXXXX   |
                            XX   |
                             X   |
                                 |
                        XXXXXX   |  4
                     XXXXXXXXX   |
                           XXX   |
                          XXXX   |
                        XXXXXX   |
                XXXXXXXXXXXXXX   |  2
                         XXXXX   |  8
   .0             XXXXXXXXXXXX   |  9
                         XXXXX   |  7 14
                     XXXXXXXXX   |  1 5 15
                          XXXX   |
                       XXXXXXX   |
                          XXXX   |
                        XXXXXX   |
                        XXXXXX   |
                       XXXXXXX   |
                          XXXX   |
                                 |
                         XXXXX   |
                            XX   |
                                 |
 -1.0                            |
                                 |
                                 |
                             X   |
                             X   |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
 -2.0                            |
Person consensus disagree   Item most difficult to endorse
-----------------------------------------------------------------------
               Each X represents 1 participant
```

Fig. 13.6 TSK items and person consensus map

map relative to their negative logit value. From Fig. 13.6, it can be seen that survey items 1 (logit −0.40), 5 and 15 are the most difficult for the participants to endorse. Conversely, Items 16 (logit +1.20), 13 (logit +1.0) and 4 (logit +0.6) are rated as the easiest for participants to endorse. It can be seen that the residual items are placed along the remainder of the hierarchical logit scale.

13.6 Discussion

Dystonia is a poorly understood neurological disorder, associated with greater depression, disability and a more negative body concept than the general population (Jahanshahi and Marsden 1990; Zetterberg et al. 2012). To establish whether kinesiophobia is a factor in the dystonia experience, this study created a dystonia-specific version of the TSK. The theoretical framework of the original TSK (Kori et al. 1990) and several aspects of validity for the new TSK-Dystonia were then explored using a Rasch analysis. In agreement with previous studies adapting the TSK to different disease populations (using different methods of analysis), we found biases affecting some items and these were removed. In the end, the 11-item TSK-Dystonia was found to represent a unidimensional construct capturing fear avoidance in people living with dystonia. We now have a valid, consistent and reliable tool to measure the construct of kinesiophobia in people with dystonia. Before implementation of the TSK-Dystonia in a clinical population, further work must be done to identify a cut-off point for when kinesiophobia is present and the minimal clinical difference to determine a positive response to interventions calculated. These steps will ensure the TSK-Dystonia is seamlessly integrated into clinical practice as an assessment tool.

The current study used a Rash analysis to test the psychometric properties of the TSK-Dystonia and to establish a unidimensional and a reliable scale. Previous studies in chronic back and neck pain, and those adapting the TSK to disorders outside of low back pain, have used methods other than Rasch analysis to assess the construct validity of their scale. These methods include confirmatory factor analysis (CFA) (Goubert et al. 2004) and principal factor analysis (PCA) (Geisser et al. 2000) and two, three and four factor models have been proposed. These constructs were considered to include fear of harm, fear of injury/re-injury, activity avoidance, a somatic focus and a tendency to catastrophise (Goubert et al. 2004; Roelofs et al. 2004; Burwinkle et al. 2005; Woby et al. 2005). It appears that at least two constructs, somatic focus and activity avoidance, are common across chronic disease populations (Goubert et al. 2004; Roelofs et al. 2004; Back et al. 2013). In contrast, a Norwegian version of TSK for chronic back pain using a Rasch analysis found the scale fitted a unidimensional construct of Kinesiophobia (Damsgard et al. 2007). The Rasch appears to be a more robust method to ensure construct unidimensionality and reliability for adaptations of the TSK scale.

In agreement with previous studies using different methods of analysis, items were removed from the original 17-item survey in the current study to ensure the TSK-Dystonia was unidimensional, allowing a valid interpretation of a score relating to kinesiophobia. There were four poorly fitting items that were removed. Poor fit usually arises because the item itself is poorly written and misinterpreted by respondents, or the survey item is sound but is not working to define the single underlying construct. Two other items were removed as were unable to differentiate between participants' consensus ability. In a study in chronic low back pain patients, item analysis led to a proposed TSK-11 (Woby et al. 2005). Another study adapting the TSK for people living with Fibromyalgia retained only four items (Burwinkle et al. 2005), while another in Temporomandibular Disorders removed 6 items (Visscher et al. 2010). The difficulties in factor models used in the past studies adapting the TSK for other disorders and diseases are that they lack dimensional and construct consensus and results are controversial (Lundberg et al. 2009). Rasch analysis allows the validation of each question separately and determines to what degree each question represents the construct. This is important for a scale like the TSK, as Likert-type scales traditionally used to measure the strength and direction of a latent variable such as attitudes are usually ordinal and not interval scales. Deriving means and standard deviations from Likert-type scales can produce misleading conclusions as with ordinal measures the categories of the Likert-type scale correctly represent an inherent order, but the numbers given to each category (e.g. 1 = very difficult, or 2 = hard, or 3 = easy, or 4 = very simple) do not indicate the magnitude of difference between the ability categories. Therefore, Likert-type scales should not be summed or averaged. However, Rasch analysis is able to transform ordinal data into true interval data and explore for disordered categories or item thresholds. The use of Cronbach alpha as a measure of reliability has limitations too, as it is unable to confirm whether the underlying constructs of the survey items measure the same underlying dimension (Shevlin et al. 2000). Rasch analysis can effectively determine rating scale reliability by identifying and removing items that do not fit with the rest of the scale and is ideal to define the underlying construct of the TSK (Lundberg et al. 2009).

One difficulty with statistical methods such as PCA analysis is that they fail to identify unpredictable responses. In previous studies, the four reversed items have been identified as unreliable and are often removed (Clark et al. 1996; Geisser et al. 2000; Woby et al. 2005). Measurement using the Rasch analysis enables identification of unpredictable responses as a quality control of the measure. Using this method, we only identified one reversed item (Item 12) that threatened scale unidimensionality. Further studies are needed to identify whether the reversed items are indeed unreliable using Rasch analysis.

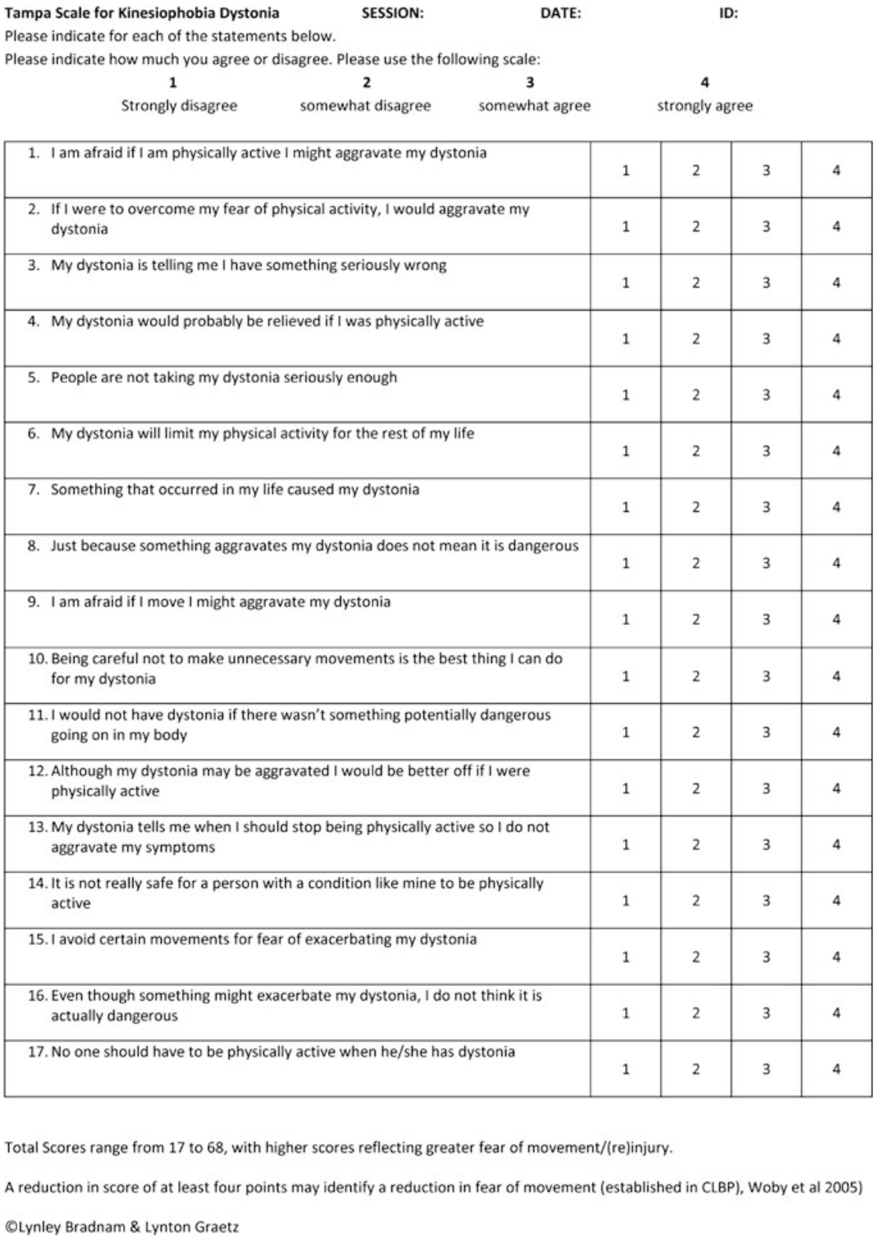
There are several limitations to this study. The first is the relatively small number of responses to the survey, particularly from males, making the impact of gender difficult to discern. However, dystonia affects women more than men, so this proportion is roughly representative of the population under study. The second is the generalizability of the study may be limited by ‘lumping’ types of dystonia other than cervical dystonia into the category of ‘other.’ Further analysis of the effect of dystonia type and gender on questions 1, 3 and 8 should be performed in a larger sample to identify whether the hint of item bias thrown up in this study is realised or not. A further concern is the estimate of consensus for the TSK-dystonia scale. The conjoint estimates of consensus plotted against the different TSK items provides a logit scale that ranges from −2 (least easiest items and least able to endorse right up to +2 which are the easier items to endorse). Ideally, the distribution of person scores should match the distribution of items scores; however, here there is a deficit in items in the lower right section of the map. This demonstrates there were insufficient survey questions to adequately measure participants' capacity to disagree with different aspects of activity avoidance in dystonia. This was particularly noticeable for items that were difficult for the participants to endorse. However, the spread of participant scores right along the scale suggests that the whole range of consensus was tested adequately. Additional question design could seek to improve the number of 'easy to endorse' items identified as found lacking by the current Rasch analysis. A final suggestion is that the scale could be changed from a consensus questionnaire to one testing self-efficacy. Self-efficacy is an important factor in patient-centred practice, as it encourages people to take active control when managing their condition (Jones et al. 2009). In this way, a range of estimates about the self-reported abilities to perform different tasks associated with dystonia could be obtained other than just consensus scores. With self-efficacy Rasch measures, items can be plotted against the raw scores and a hierarchy of self-rated abilities can be constructed to inform clinical practice interventions and targets for treatment.

13.7 Conclusion

A range of analytical methods to determine psychometric qualities have been used previous studies adapting the TSK for use in disorders besides chronic back pain with mixed results. The shortened TSK-Dystonia version might identify maladaptive beliefs about causes and effects of dystonia leading to kinesiophobia or fear avoidance in those living with the disorder. However, further adaptation of the survey is recommended so participants' consensus at the lower end of the scale (disagreement) can be more fully estimated. The influence of gender and dystonia type also needs clarifying in a larger population. Further refinement of the TSK-Dystonia could provide more questions in each category and more that are easy to for participants to agree with and endorse. The ability to determine kinesiophobia using the TSK-Dystonia scale will allow important knowledge regarding activity avoidance in people living with dystonia. The scale now should be tested to see whether it is sensitive to change with therapy and strategies that directly address movement fear avoidance in people with dystonia.

Acknowledgments The authors are grateful to the members of the expert panel for their input into survey design and all survey participants.

Appendix: The TSK-Dystonia (Original 17 Item Version)



Tampa Scale for Kinesiophobia Dystonia **SESSION:** **DATE:** **ID:**

Please indicate for each of the statements below.

Please indicate how much you agree or disagree. Please use the following scale:

1	**2**	**3**	**4**
Strongly disagree	somewhat disagree	somewhat agree	strongly agree

1. I am afraid if I am physically active I might aggravate my dystonia	1	2	3	4
2. If I were to overcome my fear of physical activity, I would aggravate my dystonia	1	2	3	4
3. My dystonia is telling me I have something seriously wrong	1	2	3	4
4. My dystonia would probably be relieved if I was physically active	1	2	3	4
5. People are not taking my dystonia seriously enough	1	2	3	4
6. My dystonia will limit my physical activity for the rest of my life	1	2	3	4
7. Something that occurred in my life caused my dystonia	1	2	3	4
8. Just because something aggravates my dystonia does not mean it is dangerous	1	2	3	4
9. I am afraid if I move I might aggravate my dystonia	1	2	3	4
10. Being careful not to make unnecessary movements is the best thing I can do for my dystonia	1	2	3	4
11. I would not have dystonia if there wasn't something potentially dangerous going on in my body	1	2	3	4
12. Although my dystonia may be aggravated I would be better off if I were physically active	1	2	3	4
13. My dystonia tells me when I should stop being physically active so I do not aggravate my symptoms	1	2	3	4
14. It is not really safe for a person with a condition like mine to be physically active	1	2	3	4
15. I avoid certain movements for fear of exacerbating my dystonia	1	2	3	4
16. Even though something might exacerbate my dystonia, I do not think it is actually dangerous	1	2	3	4
17. No one should have to be physically active when he/she has dystonia	1	2	3	4

Total Scores range from 17 to 68, with higher scores reflecting greater fear of movement/(re)injury.

A reduction in score of at least four points may identify a reduction in fear of movement (established in CLBP), Woby et al 2005)

References

Albanese, A. (1990). Extrapyramidal system, motor Ganglia and movement disorders. *Reviews in the Neurosciences, 2*(3), 145–164.

Albanese, A., Bhatia, K., Bressman, S. B., DeLong, M. R., Fahn, S., Fung, V. S., et al. (2013). Phenomenology and classification of dystonia: A consensus update. *Movement Disorders, 28* (7), 863–873.

Andrich, D. (2005). Rasch models for ordered response categories. In: *Encyclopedia of statistics in behavioral science*.

Back, M., Cider, A., Herlitz, J., Lundberg, M., & Jansson, B. (2013). The impact on kinesiophobia (fear of movement) by clinical variables for patients with coronary artery disease. *International Journal of Cardiology, 167*(2), 391–397.

Bäck, M., Jansson, B., Cider, Å., Herlitz, J., & Lundberg, M. (2012). Validation of a questionnaire to detect kinesiophobia (fear of movement) in patients with coronary artery disease. *Journal of Rehabilitation Medicine, 44*(4), 363–369.

Blackman, I., de Crespigny, C., & Parker, S. (2006). Mapping self-confidence levels of nurses in their provision of nursing care to others with alcohol and tobacco dependence, using Rasch scaling. *International Education Journal, 7*(3), 245–258.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press.

Burwinkle, T., Robinson, J. P., & Turk, D. C. (2005). Fear of movement: Factor structure of the tampa scale of kinesiophobia in patients with fibromyalgia syndrome. *J Pain, 6*(6), 384–391.

Clark, M., Kori, S., & Brockel, J. (1996). Kinesiophobia and chronic pain: Psychometric characteristics and factor analysis of the tampa scale. *American Pain Society Abstract.*

Damsgard, E., Fors, T., Anke, A., & Roe, C. (2007). The tampa scale of kinesiophobia: A Rasch analysis of its properties in subjects with low back and more widespread pain. *Journal of Rehabilitation Medicine, 39*(9), 672–678.

Geisser, M. E., Haig, A. J., & Theisen, M. E. (2000). Activity avoidance and function in persons with chronic back pain. *Journal of Occupational Rehabilitation, 10*(3), 215–227.

Goubert, L., Crombez, G., Van Damme, S., Vlaeyen, J. W., Bijttebier, P., & Roelofs, J. (2004). Confirmatory factor analysis of the tampa scale for kinesiophobia: Invariant two-factor model across low back pain patients and fibromyalgia patients. *Clinical Journal of Pain, 20*(2), 103–110.

Jahanshahi, M., & Marsden, C. (1990). Body concept, disability, and depression in patients with spasmodic torticollis. *Behavioural Neurology, 3*(2), 117–131.

Jones, F., Mandy, A., & Partridge, C. (2009). Changing self-efficacy in individuals following a first time stroke: Preliminary study of a novel self-management intervention. *Clin Rehabil, 23* (6), 522–533.

Kori, S. H., Miller, R. P., & Todd, D. D. (1990). Kinesophobia: A new view of chronic pain behaviour. *Pain Management, 3*, 35–43.

Leung, Y. -Y., Png, M. -E., Conaghan, P., & Tennant, A. (2014). A systematic literature review on the application of Rasch analysis in musculoskeletal disease—A special interest group report of OMERACT 11. *The Journal of Rheumatology, 41*(1), 159–164.

Linacre, J. (2011). A user's guide to winsteps, program manual 3.72.0. Chicago, IL: Winsteps. com.

Lundberg, M., Styf, J., & Jansson, B. (2009). On what patients does the Tampa scale for Kinesiophobia fit? *Physiotherapy Theory and Practice, 25*(7), 495–506.

Nijs, J., De Meirleir, K., & Duquet, W. (2004). Kinesiophobia in chronic fatigue syndrome: Assessment and associations with disability. *Archives of Physical Medicine and Rehabilitation, 85*(10), 1586–1592.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the hospital anxiety and depression scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1–18.

Roelofs, J., Goubert, L., Peters, M. L., Vlaeyen, J. W., & Crombez, G. (2004). The tampa scale for kinesiophobia: Further examination of psychometric properties in patients with chronic low back pain and fibromyalgia. *European Journal of Pain, 8*(5), 495–502.

Shevlin, M., Miles, J., Davies, M., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences, 28*(2), 229–237.

Tkachuk, G. A., & Harris, C. A. (2012). Psychometric properties of the tampa scale for kinesiophobia-11 (TSK-11). *The Journal of Pain, 13*(10), 970–977.

Visscher, C. M., Ohrbach, R., van Wijk, A. J., Wilkosz, M., & Naeije, M. (2010). The tampa scale for kinesiophobia for temporomandibular disorders (TSK-TMD). *Pain, 150*(3), 492–500.

Vlaeyen, J. W., & Linton, S. J. (2000). Fear-avoidance and its consequences in chronic musculoskeletal pain: A state of the art. *Pain, 85*(3), 317–332.

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models a reason for collapsing categories? *Assessment*. doi:1073191114530775.

Woby, S. R., Roach, N. K., Urmston, M., & Watson, P. J. (2005). Psychometric properties of the TSK-11: A shortened version of the tampa scale for kinesiophobia. *Pain, 117*(1–2), 137–144.

Wu, M., Adams, R., & Wilson, M. (1997). *ConQuest: Generalized item response modeling software [computer program]*. Melboune: Australian Council for Educational Research.

Yates, S. M. (2005). Rasch and attitude scales: Explanatory style. In: *Applied Rasch measurement: A book of exemplars* (pp. 207–225). Springer.

Zetterberg, L., Lindmark, B., Soderlund, A., & Asenlof, P. (2012). Self-perceived non-motor aspects of cervical dystonia and their association with disability. *Journal of Rehabilitation Medicine, 44*(11), 950–954.

Zurowski, M., McDonald, W. M., Fox, S., & Marsh, L. (2013). Psychiatric comorbidities in dystonia: Emerging concepts. *Movement Disorders, 28*(7), 914–920.

Chapter 14
Reliability and Validity of the Malay Version of the SSIS Instrument for Students Aged 8–12 Years Old in Assessing Social Skills and Problem Behavior

Zuraini Mat Issa and Wan Abdul Manan Wan Muda

Abstract The Social Skills Improvement System (SSIS) Rating Scales for students can be used to evaluate student's social skills and problem behavior. The purpose of this study was to validate and examine the reliability of the Malay version of the SSIS instrument for students aged 8–12 years old using the item analysis method. The 75-item self-administered translated SSIS instrument was introduced to 188 students from two conveniently selected primary schools in Malaysia. The polytomous data were analyzed using the Winstep version 3.80.1, which applied Rasch measurement model based on Item Response Theory (IRT) Models. The instrument was subjected to threshold calibration analysis to ensure the suitability of the scales proposed. Item reliability index was used in examining the instrument reliability, while fit statistics which include the point-measure correlation (PTMEA Corr) index and mean square (MNSQ) values and unidimensionality were examined for instrument construct validity. The results showed the proposed 4-point rating scales were workable. All SSIS subdomains have person reliability values of >0.53 and separation indexes >1.07, with positive PTMEA values for all items. The infit and outfit MNSQ that ranged between 0.5 and 1.5 were used for the purpose of reviewing and retaining items. The findings also showed that 73 were fit items (MNSQ 0.5–1.50), none were overfit and two items were misfits. Further analysis on the ICC of the misfit items suggested that those items could be retained due to careless and erratic responses. In conclusion, the Rasch measurement model could

Z.M. Issa (✉)
Department of Foodservice Management, Faculty of Hotel and Tourism Management,
Universiti Teknologi MARA, Puncak Alam Campus, 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,
16150 Kubang Kerian, Kelantan, Malaysia
e-mail: wanmanan@usm.my

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_14

be used to produce empirical evidence of validity and reliability of the translated instrument. Hence, the Malay version of the SSIS instrument for students aged 8–12 years old could then be used to further assessing student's social skills and problem behavior.

Keywords SSIS rating scales · Primary school students · Rasch measurement model · Validity · Reliability

14.1 Introduction

The intake of foods among schoolchildren has been associated with their health status (Adamsson et al. 2014; Cueto 2001), academic performance, attendance to schools (Cueto 2001; Rampersaud et al. 2005), social skills development, and problem behavior (Jyoti et al. 2005). Various methods were applied by researchers in finding associations between the intake of foods and its outcomes. The body mass index status is the most common method in assessing one's health status, while achievement in reading and math performance were evaluated in assessing student's academic performance (Shariff et al. 2000).

Similarly, various instruments were made available in determining student's social skills and problem behavior. However, these instruments are mostly for English-speaking populations and widely used in the Western countries. The instruments include Matson Evaluation of Social Skills with Youngsters (Matson et al. 1983), Social Skills Rating Systems (SSRS) Gresham and Elliott 1990), and Social Skills Improvement System (SSIS) Rating Scales (Gresham and Elliott 2008). However, the SSIS Rating Scales instrument, a revised version of Social Skills Rating System (SSRS), is the only instrument that provides comprehensive evaluation on a student.

The SSIS Rating Scales comprised three domains: (a) social skills, (b) problem behaviors, and (c) academic competence. It is the only instrument that uses a multirater approach that includes ratings from teachers, parents, and the student themselves. The teacher version comprised all the three domains, whereas parent and student forms comprised the social skills and problem behavior domains. These instruments were designed according to age-based norm. The teacher and parent versions provide norms for ages 3–5, 5–12, and 13–18, while the student version provides norms for ages 8–12 and 13–18.

To the best of our knowledge, there is neither comprehensive nor validated instrument for the Malaysian population that measures student's social skills and problem behavior. This study was performed to determine the reliability and validity of the Malay version of the SSIS Rating Scales instrument for student aged 8–12, which later would be used to measure the association between the school meal intake and their social skills and problem behavior.

14.2 Methodology

14.2.1 Research Design

This pilot study has been carried out using a quantitative survey approach in which the instrument was distributed among primary schoolchildren from two different schools in Malaysia.

14.2.2 SSIS Instrument

The 75-item SSIS (student version) instrument consists of two main domains: social skills and problem behavior. Social skills represent learning behaviors that promote positive interactions while simultaneously discouraging negative interactions. There were 46 items to determine social skills domain which were further divided into seven subdomains, that are (i) communication (7 items), (ii) cooperation (6 items), (iii) assertion (7 items), (iv) responsibility (6 items), (v) empathy (6 items), (vi) engagement (7 items), and (vii) self-control (7 items). Another 29 items were on problem behavior where these items could belong to one or more of the following four subdomains, (i) externalizing (12 items), (ii) bullying (5 items), (iii) hyperactivity/inattention (7 items), and (iv) internalizing (7 items). In addition, this instrument also has another subdomain that is an autism spectrum (15 items) which is part of problem behavior domain. However, eight and seven items were in social skills and problem behavior domains, respectively. A four-rating scale of 0 = "Not True," 1 = "A Little True," 2 = "A Lot True," and 3 = "Very True" was used in the instrument.

After obtaining approval and a license from the publisher (Pearson Assessment, Inc., USA), the SSIS (student version) instrument was translated from English to Malay language independently by a medical officer and an English teacher who were fluent in both languages. The translated versions were then revised and reconciled by a psychologist to produce a forward-translated version of the instruments. These revised instruments were then back-translated into English by another medical officer and an English teacher. Later, together with another appointed psychologist (who was also an advisor for this forward and backward translation), the researcher compared the forward- and back-translated versions and amended the instruments accordingly. The back-translated versions were then compared with the original SSIS instruments to ensure the similarity. Once researcher and the advisor agreed on the final version, the face validity was then determined.

Prior to validation and reliability analyses, the instruments were subjected to pretesting stage. In addition, due to variations in culture and dialect, it was crucial to have a pretesting stage so that the terms and expressions used would be familiar to the targeted respondents. Thus, the Malay versions of the SSIS instruments were assessed for clarity among conveniently selected groups of primary schoolchildren.

Amendments were made based on language levels and ambiguous worded items. After performing necessary corrections and amendments, the final version of the SSIS instrument for student was produced. The pretesting stage also resulted in rewording of the four-rating scale to 0 = "Not True," 1 = "Not Really True," 2 = "True," and 3 = "Very True," as the former scales used caused confusion among the respondents. The instrument was then subjected to a pilot study involving two groups of Level 2 primary schoolchildren who were conveniently identified by the respective Head Teachers from two different schools.

14.2.3 Statistical Analysis

The data were analyzed using Winsteps version 3.80.1 (SWReg, Inc., 2009) to determine the validity and reliability of the instrument based on Rasch measurement model. During the analyses, misfit persons were identified and evaluated, and whenever necessary they were removed from the data set. The demographic profile of the respondents was analyzed using SPSS version 16 (SPSS Inc., Chicago, IL) and was described using frequency, and percentage. The Rasch analysis was conducted according to

14.3 Research Findings

14.3.1 Participants

A total of 188 primary schoolchildren responded to the translated student version of the SSIS Rating Scales instrument. The demographics of the respondents are listed in Table 14.1. Most of the schoolchildren involved were female (59.6 %). The aged groups involved were mostly from aged 11 years old (48.4 %) with mean age of xx (SD).

14.3.2 Administration of the Malay Version of the SSIS Rating Scale (for Student Ages 8–12)

The instrument consists of 75 items that belong to two domains and 13 subdomains. The instrument was administered for 20–30 min with the help from the class teacher. The responses of all items are polytomous with 4-point rating scales: "Not True," "Not Really True," "True," and "Very True".

Table 1 Participant demographic characteristics (n = 188)

	Number of respondents	Percentage of respondents (%)
Gender		
Male	76	40.4
Female	110	58.5
Missing	2	1.1
Age		
9 years old	38	20.2
10 years old	14	7.4
11 years old	91	48.4
12 years old	40	21.3
Missing	5	2.7

14.3.3 Threshold Calibration

Using a Rasch analysis, the instrument met the following requirements, all of which were necessary to achieve proper rating scale function: (1) the number of observations in each category was greater than 10; (2) the average category measures increased with the rating scale categories, showing a monotonic trend; (3) infit and outfit MNSQ for measured steps were within the acceptable range (from −2 to +2 logits); and (4) category thresholds decreased with the rating scales and were at least 1.05 log odds ratio (logits) apart (Linacre 1999). Hence the proposed 4-point rating scales per original SSIS Rating Scales deemed to be used in its Malay version.

14.3.4 Reliability and Separation of Items

Table 14.2 demonstrates the Rasch analyses to provide item and person reliability and separation indexes. Item reliability indexes for all subdomains are more than 0.90, while their item separation indexes are greater than two. However, person reliability indexes in all subdomains are less than 0.80 with separation indexes all below two. According to Linacre (2007), reliability index of ≥0.8 and separation index ≥2 are good indicators for a reliable measurement scales. Nonetheless, low person reliability and separation index in measured construct indicates an insufficient number of persons that would spread along the ability continuum (Bond and Fox 2007).

Table 2 Item reliability for domains and subdomains of the SSIS Rating Scales (Malay Version for students aged 8–12)

Subdomain	Reliability		Separation		Measured Dimension (%)	Additional Dimension (%)	Item		
	Item	Person	Item	Person			Fit	Overfit	Misfit
Social skills									
Communication	0.97	0.65	5.27	1.37	45.5	1.7	6	1	–
Cooperation	0.98	0.63	7.40	1.32	57.0	1.6	6	–	–
Assertion	0.96	0.59	5.17	1.20	44.6	1.7	7	–	–
Responsibility	0.96	0.56	5.07	1.14	49.0	2.0	6	–	–
Empathy	0.96	0.62	4.82	1.27	45.0	1.7	6	–	–
Engagement	0.95	0.71	4.52	1.57	54.6	1.8	5	1	1
Self-Control	0.97	0.76	5.73	1.78	56.7	1.5	6	1	–
Problem behaviour									
Externalizing	0.95	0.76	4.38	1.80	41.0	1.9	11	1	–
Bullying	0.92	0.71	4.45	1.56	49.8	1.9	4	–	1
Hyperactivity/Inattention	0.95	0.79	4.60	1.96	40.9	1.5	6	1	–
Internalizing	0.96	0.80	5.11	2.01	51.6	1.4	6	–	–
Autism spectrum (social skills)	0.95	0.64	4.60	1.34	42.7	1.8	7	1	–
Autism spectrum (problem behavior)	0.96	0.58	4.96	1.17	47.4	1.6	7	–	–

14.3.5 Unidimensionality

It is expected that instrument used to be unidimensional and should provide different levels of difficulty and fair to all respondents. The translated instrument used in this study was subjected to the principal component of analysis (PCA) for the standardized residuals for each subdomain. The PCA output for each subdomain is shown in Table 14.2. All subdomains except for Communication and Autism Spectrum (Social Skills) have strong measured variances of more than 40 % (Linacre 2006). However, all subdomains have an additional dimension variance less than 3 % (Bond and Fox 2007).

14.3.6 Fit Statistics

In Rasch analysis, item fit statistics can be used to evaluate the extent to which items are tapped into the same construct and places test takers in the same order to assess the item's technical quality empirically (Smith 1992). Therefore, statistical analysis for items were carried out suitability to identify items that have positive point-measure correlation (PTMEA Corr), the infit and outfit MNSQ values greater than 0.5 and less than 1.5, and the Z-standardized (ZSTD) values between −2.00 and 2.00. The item was considered misfit to the model if the item has the infit and/or outfit MNSQ value exceeding the range of 1.5 (Linacre 2006). A fit statistics >1.5 means that the item may not contribute the same underlying construct as do the other items in the same scale. An item with a fit statistics <0.5 indicates redundancy of the items in the same scale (Duncan et al. 2003). Conclusion was made after the fit statistics was performed on the PTMEA Corr, outfit MNSQ followed by the infit MNSQ, outfit ZSTD, and infit ZSTD (Bond and Fox 2007).

As shown in Table 14.2, all items in Cooperation, Assertion, Responsibility, Empathy, Internalizing, and Autism Spectrum (Problem Behaviour) subdomains fit the model. However, six items from six different subdomains were considered as overfit and another two items which belong to Engagement (1 item) and Bullying (1 item) subdomains were regarded as misfit items.

14.4 Conclusion

In general, the findings of this study indicated that the content of 75 items in the Malay version of SSIS Rating Scale (for student aged 8–12) is valid and reliable. The items in each subdomain were shown to have high reliability indexes, despite poor reliability and separation indexes for person. High item reliability indicates the stability of the item. It is further strengthened with strong item separation index that exceeded the acceptable value of 2.00.

In addition, all subdomains except for Communication and Autism Spectrum (Social skills) have strong measurement dimension (>40 %), and the primary PCA analysis was further supported by additional measurement dimension of less than

3 %. Fit statistics outputs were used to evaluate construct validity of all subdomains. The outputs revealed that only two items from the Malay version were misfit and should be removed from the instrument. However, further analysis on the ICC had resulted that the items could be retained.

The Rasch measurement model based on IRT model was shown to be effective in producing a valid and reliable translated instrument. Future study is recommended to focus at differences based on GDIF items to remove the bias item based on gender and level of students.

Acknowledgment The authors would like to express their utmost appreciation to the Ministry of Education (MoE) and Universiti Teknologi MARA, Malaysia for providing support and financial assistance through the Higher Learning Institution (HLI) Academic Training Scheme (SLAI). The authors would also like to express our gratitude to the school teachers and students for their support in making this research project possible.

References

Adamsson, V., Reumark, A., Marklund, M., Larsson, A., & Risérus, U. (2014). Role of a prudent breakfast in improving cardiometabolic risk factors in subjects with hypercholesterolemia: A randomized controlled trial. *Clinical Nutrition (Edinburgh, Scotland)*. doi:10.1016/j.clnu.2014.04.009

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

Cueto, S. (2001). Breakfast and performance. *Public Health Nutrition, 4*(6a), 1429–1431. doi:10.1079/PHN2001233.

Duncan, P. W., Bode, R. K., Min Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Archives Physics Medicine Rehabilitation*, *84*(7), 950–963

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system: American guidance service*. Minneapolis: Circle pines.

Gresham, F. M., & Elliott, S. N. (2008). *SSIS rating scales manual*. Minneapolis: Pearson.

Jyoti, D. F., Frongillo, E. A., Jones, S. J., & Al, J. E. T. (2005). Food. *The Journal of Nutrition*, *135*(May 2005), 2831–2839.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103–122.

Linacre, J. M. (2006). Data variance explained by measures. *Rasch Measurement Transactions, 20* (1), 1045.

Linacre, J. M. (2007). *A user's guide to WINSTEP Rasch—model computer programs*. Chigago: MESA Press.

Matson, J. L., Rotatori, A. F., & Helsel, W. J. (1983). Development of a rating scale to measure social skills in children: The Matson Evaluation of Social Skills with Youngsters (MESSY). *Behavioral Research and Theory, 21*, 335–340.

Rampersaud, G. C., Pereira, M. a, Girard, B. L., Adams, J., & Metzl, J. D. (2005). Breakfast habits, nutritional status, body weight, and academic performance in children and adolescents. *Journal of the American Dietetic Association*, *105*(5), 743–60; quiz 761–762. doi:10.1016/j.jada.2005.02.007

Shariff, Z. M., Bond, J. T., & Johnson, N. E. (2000). *Nutrition and Educational Achievement of Urban Primary Schoolchildren in Malaysia, 9*(June), 264–273.

Smith, R. M. (1992). *Applications of rasch measurement*. Chigago: MESA Press.

Chapter 15
Sample Size and Chi-Squared Test of Fit—A Comparison Between a Random Sample Approach and a Chi-Square Value Adjustment Method Using Swedish Adolescent Data

Daniel Bergh

Abstract *Background* Significance tests are commonly sensitive to sample size, and Chi-Squared statistics is not an exception. Nevertheless, Chi-Squared statistics are commonly used for test of fit of measurement models. Thus, for analysts working with very large (or very small) sample sizes, this may require particular attention. However, several different approaches to handle a large sample size in test-of-fit analysis have been developed. Thus, one strategy may be to adjust the fit statistic to correspond to an equivalent sample of different size. This strategy has been implemented in the RUMM2030 software. Another strategy may be to adopt a random sample approach. *Aims* The RUMM2030 Chi-Square value adjustment facility has been available for a long time, but still there are few studies describing the empirical consequences of adjusting the sample to correspond to a smaller effective sample size in the statistical analysis of fit. Alternatively, a random sample approach could be adopted in order to handle the large sample size problem. The purpose of this study was to analyze and compare these two strategies as test-of-fit approximations, using Swedish adolescent data. *Sample* The analysis is based on the survey Young in Värmland which is a paper-and-pencil-based survey conducted recurrently since 1988, targeting all adolescent in school-year 9 residing the county of Värmland, Sweden. So far, more than 20,000 individuals have participated in the survey. In the analysis presented here, seven items based on the adolescents, experiences of the school environment were subjected to analysis, in total 21,088 individuals. *Methods* For the purposes of this study, the original sample size was adjusted to several different effective samples using the RUMM2030 adjustment function, in the test-of-fit analysis. In addition, 10 random samples for each sample size were drawn from the original sample and averaged Chi-Square values calculated. The Chi-Square values obtained using the two strategies were compared.

D. Bergh (✉)
Centre for Research on Child and Adolescent Mental Health, Karlstad University,
651 88 Karlstad, Sweden
e-mail: daniel.bergh@kau.se

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_15

Results Given the original sample of 21,000, adjusting to samples 5,000 or larger, the RUMM2030 adjustment facility work as well as a random sample approach. In contrast, when adjusting to lower samples, the adjustment function is less effective in approximating the Chi-Square value for an actual random sample of the relevant size. Hence, fit is exaggerated and misfit underestimated using the adjustment function, in particular that is true for fitting but not misfitting items. *Conclusion* Although the inferences based on p values may be the same, despite big Chi-Square value differences between the two approaches, the danger of using fit statistics mechanically cannot be enough stressed. Neither the adjustment function nor the random sample approach is sufficient in evaluating model fit; instead, several complementing methods should be used.

Keywords Polytomousrasch model · Rasch model for ordered response categories · Fit analysis · Chi-Squared statistics · Random samples · Adjusted samples

15.1 Introduction

Significance tests are commonly used in order to evaluate the fit of measurement models, e.g., measurement models based on the Rasch model. Thus, significance tests (e.g., Chi-Squared) are used in order to study the concordance between the data and the expected Rasch model. However, significance tests are also sensitive to sample size, implying that using a large enough sample also (proportionally) *trivial* differences will turn up as *statistically* significant (Martin-Löf 1973, 1974). The opposite problem, the case in which a too small sample is subjected to analysis implying that it is not possible to *statistically* identify substantial differences, should also be recognized. At the same time, a one critical number solution applicable for most situations for reporting how well the data fits the model is often preferred (Smith et al. 1998), not least by journal editors.

The combination of using large samples and relying on *p values* or *test values* mechanically as the only source of information in evaluating the concordance between the data and the model may lead to false conclusions, a phenomenon sometimes labeled "The large sample size fallacy" (Lantz 2013).

From a Rasch measurement perspective, the implications of the "The large sample size fallacy" implies an opposite problem compared to in general quantitative analysis in the social sciences. Statistics, for instance Chi-Squared, are commonly used in order to analyze the concordance between the data and the expected Rasch model (Rasch 1960). Thus, when using a large sample size, the parameters will be estimated with great precision, which further means that even very small differences between the expected Rasch model and the observed data will be readily exposed, and consequently, no items are likely to fit the model (Andrich 1988). Put differently, when applying a large sample, the power to detect misfit is so great that even if observed and expected values are very close, all items will misfit (Andrich et al. 2009). Therefore, using a large sample and mechanically relying on traditional fit statistics will almost automatically reject any model tested.

In this study, the concordance between observed data and the expected Rasch model is analyzed by means of the Rasch model for ordered response categories, also called the polytomous Rasch model, which is an extension of the Simple Logistic Model:

$$\Pr\{X_{ni} = x\} = \frac{e^{x\beta_n - \delta_i)}}{1 + e^{\beta_n - \delta_i}}.$$

Thus, in the dichotomous case, item locations are denoted by δ and person locations denoted by β. The relationship between items and persons is central; the probability of a specific response category is a function of the relationship between person parameter estimates and item parameter estimates, consequently $\beta - \delta$. A positive value from the subtraction implies probabilities greater than 0.5. Commonly, social scientific data are not restricted to dichotomous response formats. Instead, the polytomous response format may be more applicable. The Rasch model for ordered response categories (Andrich 1978; Wright and Masters 1982) takes the general form:

$$\Pr\{x_{ni} = x\} = \frac{e^{-\tau_{1i} - \tau_{2i} \cdots - \tau_{xi} + x(\beta_n - \delta_i)}}{\sum\limits_{x'=0}^{m_i} e^{-\tau_{1i} - \tau_{2i} \cdots - \tau_{x'i} + x'(\beta_n - \delta_i)}}.$$

Thus, in the polytomous case, a central concept is threshold. Given a situation with five response categories (0, 1, 2, 3, and 4), a threshold specifies the point at which the probability for choosing one of two answers is equal, for instance, an answer of 0 or 1. In the equation mentioned earlier, the threshold parameter is denoted by τ and the item score by x in the numerator.

In order to be able to depict the data as well as possible, a suggested strategy is to use several sources of information, including formal statistical analysis of fit as well as descriptive graphical analysis (Andrich 1988). This would also make it easier to handle very large or very small samples in the analysis of fit. Unfortunately, the journal format often implies limited space, which means that it is not always possible to include graphical or analyses by other means that require much space. However, several different approaches in order to handle the sample size problem have been discussed elsewhere (see for instance: Gustafsson 1980; Smith et al. 1998; Tennant and Pallant 2012; Wright and Masters 1982; Wright and Linacre 1994; Wright and Masters 1990).

In the RUMM2030 program (Andrich et al. 2013), a facility enabling sample size adjustment in the statistical analysis of fit has been implemented, implying that the Chi-Square value, all other things being equal, is adjusted in the analysis. The Chi-Square statistic for test of fit is conducted by comparing the total score of persons in approximately equal-sized class intervals, with the sum of expected values. This is resulting in an approximate Chi-Square statistic with $C - 1$ degrees of freedom and which is, in accordance with Andrich and Styles (2011 PG 81), denoted by:

$$X^2_{C-1,i} \approx \sum_{c=1}^{C} \left(\left(\sum_{n \in c} X_{ni} - \sum_{n \in c} E[X_{ni}] \right)^2 \Big/ \sum_{n \in c} V[X_{ni}] \right).$$

In order to adjust the analysis to a smaller equivalent sample size (n), the Chi-Square value obtained using the original sample size (N) is suggested to be multiplied by *n/N* (Andrich and Styles 2011). Thus, by adjusting the analysis of fit to a smaller equivalent effective sample, the Chi-Square test of fit may be considered to be less sensitive to sample size but with the residuals reflecting the degree of precision available in the original sample (Andrich and Styles 2011).

15.1.1 Aims

The adjustment facility has been implemented in the RUMM2030 program for a long time, but still there seems to be a lack of studies describing the empirical consequences of adjusting the statistical test of fit to a sample of different effective size. As far as I know, only one article addressing the operational characteristics of the adjustment function have been published so far (Bergh 2015), focusing on overall fit using simulated data. However, no paper has addressed the problem using real data. An alternative to adjusting the sample size in order to handle the sample size problem may be to utilize a random sample approach. The purpose of this study was to analyze and compare these two strategies as test-of-fit approximations using Swedish adolescent data.

15.2 Methods

15.2.1 Material

This study uses data from the paper-and-pencil survey Young in Värmland, conducted among Swedish adolescents in school-year 9, in the compulsory Swedish school. This means that the respondents are of 15–16 years of age. Young in Värmland aims to monitor the health and well-being of all adolescents in school-year 9 in the county of Värmland and have been conducted recurrently since 1988. So far, more than 20,000 students have participated in the survey. The questionnaire includes item topics about the school, school work, activities during after school hours, bullying and violence, worry and self-esteem, exercise and diet, and the use of alcohol and tobacco as well as illicit drugs.

15.2.2 Data Collection

The data collection took place during the second semester of each year of investigation. The paper-and-pencil questionnaire was distributed by school personnel and completed anonymously in the classroom. After completion, the student returned the questionnaire in a sealed envelope to responsible school personnel. The students were informed about the voluntary nature of participation. The principles guiding the data collection were approved by the ethical committee at Karlstad University, from 1995 onward. Data are collected from students who resided in the 16 municipalities within the County of Värmland. However, only data from the municipalities participating all years of investigation are subjected to analysis (14 of 16). The number of students at each year of investigation was 3202 (1988), 3049 (1991), 2435 (1995), 2741 (1998), 2931 (2002), 3124 (2005), 3109 (2008), 2620 (2011). The following attrition rates apply: 10.0 % (1988), 10.9 % (1991), 6.3 % (1995), 9.4 % (1998), 11.3 % (2002), 13.9 % (2005), 15.7 % (2008), and 16.6 % (2011). Due to extreme scores, the estimation procedure excluded some individuals. In total 21,088 individuals were subjected to analysis.

15.2.3 Instrument

In this particular analysis, seven items on students' experiences of their learning environment is used. However, as the focus in this paper is not on the specific items but on fit analysis as such, and the characteristics of individual items are not reported in detail. The students had to answer seven items on what they believe characterizes the school work in their class and with a five-point response format ("Never," "Seldom," "Sometimes," "Often," and "Always"). The wording of the items implies that a low score means that the students experiences the learning environment as poor, while a high score implies positive experiences of the school class as a learning environment.

15.2.4 Procedure

15.2.4.1 Adjustment

For the purposes of this study, the original sample was adjusted to several different effective sample sizes using the RUMM2030 adjustment function. This facility has been described in detail by Andrich and Styles (2011). Here, it is just concluded that in order to adjust the Chi-Square value from analyses based on the full original sample (*N*) to a smaller equivalent sample (*n*), the original Chi-Square value should be multiplied by *n/N*, as suggested by Andrich and Styles (2011). For facilitating the study of Chi-Square value change as moving from one sample size to a smaller, the original sample was adjusted to several different samples of different size. Thus, the original sample of 21,088 individuals was adjusted to 10000; 7000; 5000; 3000; 2000; 1000; 750; 500; 300; and 100 individuals. For each sample size, traditional fit statistics in the form of individual item Chi-Square values and their corresponding Mean-Squares as well as total (overall) Chi-Square values were calculated. Mean-Square values were calculated by dividing the Chi-Square values with their degrees of freedom (*df*), resulting in an expected value of 1.

15.2.4.2 Random Samples

In order to facilitate the comparison of the Chi-Square values obtained using the adjustment function with external values, i.e., for Chi-Square values obtained not using the adjustment function, a random sample approach was adopted. From a statistical point of view, averaged Chi-Square values based on sets of random samples may be considered to be a good Chi-Square value approximation. Thus, 10 random samples with replacement were drawn from the original sample and averaged total Chi-Square values calculated. In addition, Mean-Square values were calculated by dividing the Chi-Squares with their *df*, resulting in an expected value of 1. Thus, well-fitted data should have a Mean-Square value close to one, and misfitted data imply Means-Squares substantially above or below the value of 1. It should, however, be noticed that the generation of random samples of similar size as the original one would imply that they will be very dependent on each other, i.e., containing the same individuals. Therefore, the random sample and adjustment procedures includes about half (10,000) of the original sample size (21,000).

15.3 Results

15.3.1 Targeting

In Table 15.1, the proportion in different response categories is displayed for each of the seven items subjected to analysis. From Table 15.1, it is evident that a smaller proportion has replied in one of the categories "Never" or "Seldom" than in one of the categories "Often" or "Always." This response pattern indicates that students generally have positive perceptions of the school class as a learning environment. In a somewhat more detailed analysis, it can be seen that a very small proportion of the respondents have chosen the response alternative "Never" (0), but a larger proportion replies in the "Seldom" (1) category while 60–80 % of the subjects reply in one of the categories: "Sometimes" (2) or "Often" (3). Finally, the "Always" (4) alternative is more common than the alternatives "Never" (0) or "Seldom" (1).

Table 15.1 The proportion in different response categories (n = 21,088)

Item	Response category				
	Never (0)	Seldom (1)	Sometimes (2)	Often (3)	Always (4)
Item 1	1	6	40	46	7
Item 2	4	12	36	39	8
Item 3	1	8	23	45	22
Item 4	2	9	25	46	18
Item 5	9	15	22	29	24
Item 6	8	22	32	25	12
Item 7	3	7	23	37	30

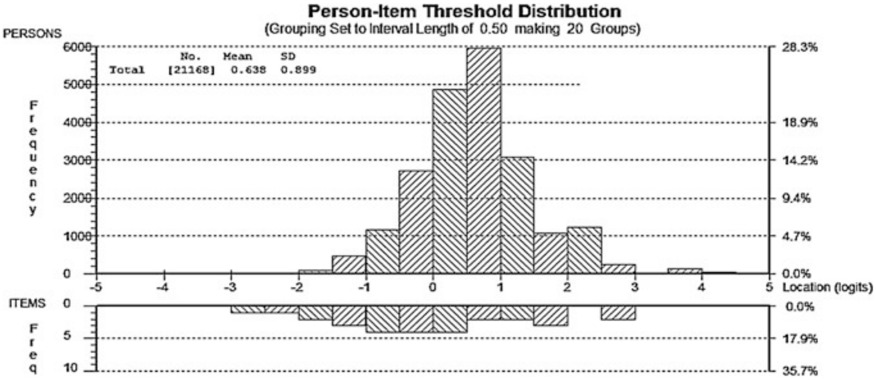


Fig. 15.1 Person-item threshold distribution

Figure 15.1 shows the distribution of persons relative to the item thresholds on their common latent trait. A low location value (logit) implies negative perceptions of the school class as a learning environment (to the left on the latent trait), while a higher score implies more positive perceptions of the same learning environment. The distribution is negatively skewed, with a positive mean (0.638), confirming the pattern observed in Table 15.1, i.e., most students have positive experiences of school as a learning environment. It is also evident that there are item thresholds where most of the persons are located. This is important since these are the locations where most information is available, i.e., information is available where the people are located. However, at locations just above 2 logits, there seems to be a tendency of a potential lack of information, as there are no item thresholds at this location. It can also be seen that there are relatively more item thresholds available in the beginning of the continuum compared to in the end, i.e., there are more information available for students experiencing poor school environments compared to those experiencing better ones.

Table 15.2 shows test-of-fit analysis for each of the individual items as well as an overall test of fit using the original sample of 21,088 individuals. Evidently, all

Table 15.2 Test of fit

Item	Location	Chi-Square	Probability	Fit residual	Mean-Square
Item 1	−0.241	64.25	0.000000	10.066	7.138333
Item 2	0.313	39.66	0.000009	6.473	4.406222
Item 3	−0.401	116.68	0.000000	−0.991	12.975556
Item 4	−0.168	28.06	0.000932	4.244	3.117778
Item 5	0.268	81.82	0.000000	−3.503	9.091111
Item 6	0.490	266.62	0.000000	−9.287	29.624444
Item 7	−0.260	85.28	0.000000	−0.404	9.475556
Total Chi-Square		682.36	0.000000		10.8

Individual item Chi-Square and total (overall) Chi-Square using the original sample (n = 21,088), with persons divided into 10 class intervals (*df* individual items = 9, and *df* = 63 for total Chi-Square

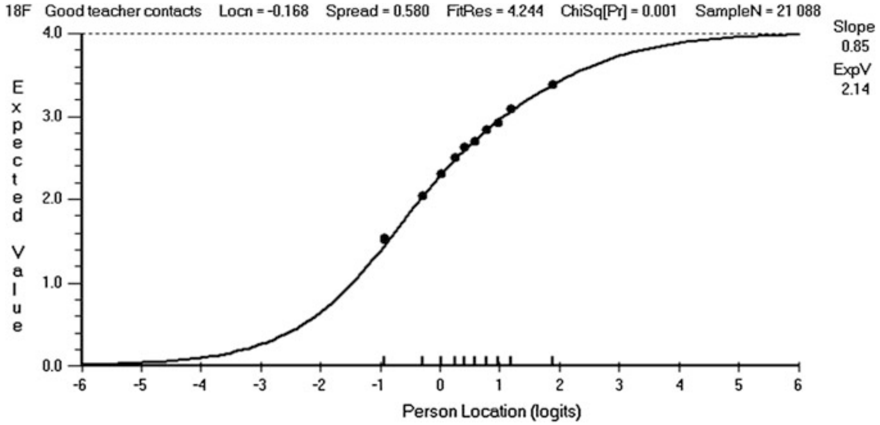


Fig. 15.2 Item characteristic curve (ICC) for Item 4, the best fitting item using the original sample

items show misfit according to the Chi-Squared statistic using the original sample. In particular, Item 6 is misfitting, showing a Mean-Square value almost 30 times higher than the expected value of 1. However, it is also possible to observe an overall misfit, also reflected by a Mean-Square value of 10.8, i.e., almost 11 times higher than the expected value of 1. However, also Item 4 which is the best-fitting item shows misfit using the original sample size. Thus, using an average of 10 random samples, it was possible to identify the best- and worst-fitting items from the original sample. However, it was not possible to depict the complete item order using the random sample method.

Figure 15.2 shows the graphical representation of the above-described test of fit for the best-fitting item 4 by means of an Item Characteristic Curve (ICC). Thus, despite a significant Chi-Square value, and a Mean-Square value substantially different from the expected value of 1, all the observations are located on or very close to the expected curve according to the Rasch model. Thus, graphically, this item seems to fit the model well.

In Table 15.3, the total Chi-Square values obtained using the RUMM2030 adjustment facility is compared with those obtained using a random sample approach. Generally, moving from one sample size to a smaller implies lower Chi-Square values which is true for adjusted samples as well as random samples. Another consistent pattern is that the random sample–adjusted sample ratio is greater than 1, i.e., the Chi-Square value obtained using the random sample approach is larger than that obtained using the adjustment function. Generally, the ratio is increasing while moving from one sample size to a smaller. However, between sample sizes of 10,000 and 7,000, it is not possible to observe a increase in ratio. The differences between the two approaches are smaller at larger samples, implying that the random sample approach and the adjustment function seem to work in a similar manner when adjusting to these sample sizes, given the original sample of 21,088. However, when adjusting to smaller samples than 2,000, the differences are increasing substantially, i.e., the methods are not comparable as test-of-fit approximations. This is also reflected in the Mean-Square values. For instance, adjusting the sample to 100 individuals implies a Mean-Square value 20 times as low as the expected value of 1. However, despite large differences in total Chi-Square values, using a sample of 1,000 individuals, the analyst would reach the same conclusion based on *p values* on how whether the data fits the model.

Table 15.4 shows the test of fit for individual items using, the original sample, an adjusted sample to 1,000 individuals and an average of Chi-Square values from 10 random samples of 1,000. It is important to recognize that the item order (according to the level of fit) is the same in the original sample and the adjusted sample. However, it is also important to realize that the best- and worst-fitting items, Item 4 and Item 6, respectively, are the same in all three cases but that the order among some other items is different using random samples due to random variation.

Table 15.5 shows comparisons between the two methods for Item 4 and Item 6, the best- and worst-fitting items, respectively. Again, using the adjustment function results in smaller Chi-Square values, i.e., better fit compared to using a random sample approach. In particular for the best-fitting item, the differences are pronounced. With adjustments between 10,000 and 3,000 individuals, given the original sample of 21,088, only small differences can be observed, i.e., the random sample–adjusted sample ratio is small. However, when adjusting to samples of 2,000 or smaller, the two methods are not comparable, i.e., the random sample–adjusted sample ratio is increasing substantially and gradually. However, when analyzing the worst-fitting item, it is evident that the differences between the two approaches are smaller. For instance, adjusting to samples between 10,000 and 500 individuals, only small differences are observable, i.e., the two methods are comparable. Only when adjustments down to 300 or 100 individuals, given the original sample of 21,088, are conducted, it is possible to observe substantial differences between the Chi-Square values obtained using the random sample approach and the adjustment function. The differences are further illustrated graphically in Fig. 15.3.

Table 15.3 Comparisons between total Chi-Square values based on different sample sizes using the RUMM2030 adjustment function, and average total Chi-Square values based on 10 random samples for each sample size

Sample size	Chi-Square original sample	Prob.	Adjusted Chi-Square	Prob. adjusted Chi-Square	Mean-Square adjusted Chi-Square	Averaged Chi-Square random samples	Prob. random samples	Mean-Square random samples	Ratio averaged Chi-Square/adjusted Chi-Square
21,088	682.36	0.0000							
10,000			323.6	0.0000	5.14	346.9	0.0000	5.51	1.1
7,000			226.5	0.0000	3.59	254.8	0.0000	4.05	1.1
5,000			161.7	0.0000	2.57	199.7	0.0000	3.17	1.2
3,000			97.1	0.0038	1.54	135.6	0.0000	2.15	1.4
2,000			64.7	0.4164	1.03	105.9	0.0006	1.68	1.6
1,000			32.4	0.9995	0.51	82.9	0.0469	1.32	2.6
750			24.3	0.9997	0.39	71.3	0.2212	1.13	2.9
500			16.2	1.0000	0.26	55.6	0.7356	0.88	3.4
300			9.7	1.0000	0.15	61.9	0.5170	0.98	6.5
100			3.2	1.0000	0.05	54.6	0.7656	0.87	17.4

The corresponding Mean-Square values also provided (7 polytomous items, 10 class intervals, *df* = 63)

Table 15.4 Individual item test of fit

Item	Chi-Square original sample	Probability original sample	Chi-Square adjusted sample to 1000	Probability adjusted sample to 1000	Chi-Square average of random samples of 1000	Probability average of random samples of 1000
Item 1	64.25	0.00	3.05	0.96	9.88	0.36
Item 2	39.66	0.00	1.88	0.99	10.25	0.33
Item 3	116.68	0.00	5.53	0.61	14.74	0.01
Item 4	**28.06**	**0.00**	**1.33**	**0.99**	**7.68**	**0.57**
Item 5	81.82	0.00	3.88	0.43	12.39	0.19
Item 6	**266.62**	**0.00**	**12.64**	**0.18**	**17.84**	**0.04**
Item 7	85.28	0.00	4.04	0.91	10.15	0.34

Original sample of 21,088, adjusted sample to 1000, and an average of 10 random samples of 1000 (7 polytomous items, 10 class intervals, *df* = 9), best and worst fitting items bolded

15.4 Discussion

The purpose of this paper was to analyze and compare two methods for handling the sample size problem in traditional test-of-fit analysis. Thus, the adjust sample size function, implemented in the RUMM2030 software, was compared to an approach utilizing random samples in order to be able to use test-of-fit statistics. This paper also further advances analyses conducted in a recently accepted paper using simulated data (Bergh 2015). In order to address the sample size problem using real data and with a more pronounced focus on individual items, the present paper complements previous work.

Focusing on the overall test of fit, the analysis presented here shows that the adjust sample size function and an average of random samples both provide similar results when adjusting the original sample (N = 21,088) to larger samples than 2,000 individuals. However, when adjusting to smaller samples, the random sample approach and the adjust sample size function shows divergent results, implying that the Chi-Square values obtained using the two methods are not fully comparable in these circumstances. Despite this, the analyst may reach the same conclusion on whether the latent trait fits the Rasch model or not, focusing on *p values* rather than the Chi-Square values as such, which also confirms the results from the above-mentioned paper (Bergh 2015).

In accordance with previous research (for instance, Martin-Löf 1973, 1974), using the original sample results in overall misfit but also that all individual items show significant misfit, despite very small substantial deviations between the expected Rasch model and the data (see example with Item 4). Using the adjustment function, item order is maintained as in the original sample, i.e., items are ordered by means of how well they fit the Rasch model. However, using the random sample approach, it was also possible to identify the best- and worst-fitting

Table 15.5 Comparisons between individual item Chi-Square values based on different sample sizes using the RUMM2030 adjustment function, and average Chi-Square values based on 10 random samples for each sample size for the best and worst fitting items (Item 4 and Item 6), *df* = 9

Sample size	Chi-Square original sample	Adjusted item Chi-Square value	Averaged Chi-Square value from random samples	Ratio averaged Chi-Square value/adjusted Chi-Square value	Chi-Square original sample	Adjusted Chi-Square value	Averaged Chi-Square value from random samples	Ratio averaged Chi-Square value/adjusted Chi-Square value
	Best fitting item				Worst fitting item			
21.088	28.06				266.62			
10,000		13.3	15.7	1.2		126.4	132.3	1.1
7,000		9.3	14.9	1.6		88.5	87.0	1.0
5,000		6.7	14.7	2.2		63.2	67.4	1.1
3,000		3.9	8.9	2.3		37.9	40.8	1.1
2,000		2.7	8.5	3.1		25.3	27.2	1.1
1,000		1.3	8.6	6.6		12.6	17.8	1.4
750		0.9	7.2	8.0		9.5	15.6	1.6
500		0.6	6.2	10.3		6.3	8.8	1.4
300		0.4	6.2	15.5		3.8	9.9	2.6
100		0.13	7.9	60.8		1.3	7.3	5.6

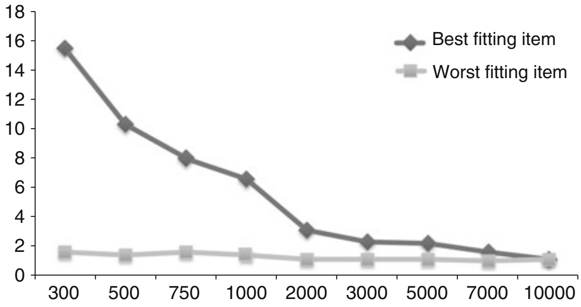


Fig. 15.3 The random sample/adjustment function ratio at different sample sizes for the best and worst fitting items

items, i.e., they were the same as in the original sample and in the adjusted sample. In this study, the best- and worst-fitting items were analyzed in some detail. Interestingly, the differences between the two approaches are most pronounced for the best-fitting item. In particular, adjusting the sample to sample sizes smaller than 2,000 individuals, given the original sample of 21,088, the differences are highly pronounced. Nevertheless, the analyst would reach the same conclusion on whether the data fits the Rasch model or not focusing on *p values* rather than the differences between Chi-Square values. Analyzing the worst-fitting item, it is evident that only small differences between the two methods appear in these circumstances, i.e., the Chi-Square values obtained using the two methods are comparable.

The analyses presented in this paper show that the adjustment facility implemented in RUMM2030 result in exaggerated Chi-Square values and their corresponding *p values* for the overall test of fit as well as for the best-fitting item. Thus, adjusting to a smaller sample size than 2,000 individuals reveal Chi-Square values much smaller than the *df* and p values approaching 1.0. Also, the corresponding Mean-Square values are much lower than the expected value of 1, which by some scholars would be interpreted as overfit (Linacre, 2002), suggesting that the observations are too predictable. However, it should not be surprising to obtain these exaggerated Chi-Square values in these circumstances. When adjusting the Chi-Square value from a large sample to correspond to a much smaller effective sample, the level of precision and the residuals available are the same as in the much larger sample, but with a much lower power to detect misfit, the exaggerated Chi-Squares are, therefore, expected.

However, using a random sample approach, the results reported here revealed that although the best- and worst-fitting items were the same, the item order was not completely the same as in the original sample. Therefore, using only 10 random samples is not considered to be sufficient in order to map the characteristics of the original sample. Therefore, more random samples would be needed if they are to be used as test-of-fit approximations.

Thus, as the adjustment facility is providing exaggerated results (with overfitting data) when adjusting to a small sample given a large original sample size, and as the few random samples used here is not able to reflect the characteristics of the original sample, none of the two methods for test-of-fit approximations are to be considered as sufficient. Instead, they are both providing important information on how well the data fit the Rasch model.

However, the adjustment facility could be used in order to estimate the accuracy of averaged Chi-Square values obtained using sets of random samples, i.e., how well they are sufficient in order to characterize the original sample. Ideally, in the future, software developers will provide the facility of generating numerous random samples and to calculate averaged fit statistics based on these. Nevertheless, fit statistics is only one piece of information necessary in order to study the concordance between the data and the expected model, also other means of fit analyses should be undertaken. Nevertheless, it is sometimes a requirement from journal editors to provide the level of fit with a one number solution. Therefore, it may be important to find a method for handling very large or very small samples, in the test-of-fit analysis.

15.5 Conclusion

In this paper, two methods to handle large samples in test-of-fit analysis were compared: the adjust sample size procedure available in the RUMM2030 software and a strategy using random samples. Using the adjust sample size function, given an original sample of 21,000, conducting adjustments down to the order of 2,000 or lower there is a risk of exaggerating fit and underestimating misfit. Therefore, in these circumstances, the adjustment function is not considered to be effective at approximating the Chi-Square value of an actual random sample of relevant size, i.e., the Chi-Square values obtained using the two methods are not comparable. The differences are larger for fitting items than for misfitting items. Using fit statistics and adjusting samples mechanically may lead analysts to spuriously accepting misfitting data. Nevertheless, when adjusting to very small samples, given the original sample of 21,000, the inferences based on p values may be the same, despite big Chi-Square value differences between the two approaches. Neither the adjustment function nor the random sample approach is sufficient in evaluating model fit, instead several complementing methods should be used. Working with large samples, the adjustment function may be used as a heuristic tool in the analysis of fit. By adjusting the sample to a smaller effective sample, it is possible to identify the relationship between items in terms of best- and worst-fitting items, while the actual level of fit may be estimated using a random sample approach. Thus, the adjustment function may serve as a tool in determining whether the random samples are accurate or not.

References

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

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage Publications.

Andrich, D., Sheridan, B., & Luo, G. (2009). *Interpreting RUMM2030 (Part I, Dichotomous Data): Rasch unidimensional models for measurement*. Perth, Western Australia: RUMM Laboratory Pty Ltd.

Andrich, D., Sheridan, B., & Luo, G. (2013). *RUMM2030: A windows program for the rasch unidimensional measurement model [Computer Software]*. Perth, WA, Australia: RUMM Laboratory.

Andrich, D., & Styles, I. (2011). Distractors with information in multiple choice items: A rationale based on the Rasch model. *Journal of Applied Measurement, 12*, 67–95.

Bergh, D. (2015). Chi-squared test of fit and sample size—A comparison between a random sample approach and a chi-square value adjustment method. *Journal of Applied Measurement*, Forthcoming.

Gustafsson, J.-E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33*, 205–233.

Lantz, B. (2013). The large sample size fallacy. *Scandinavian Journal of Caring Sciences, 27*, 487–492.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*, 878.

Martin-Löf, P. (1973, May 7–12). *The notion of redundancy and its use as a quantitative measure of the deviation between a statistical hypothesis and a set of observational data*. Paper presented at the Conference on foundational questions in statistical inference, Aarhus, Denmark.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*, 3–18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Denmark, Copenhagen, Danish Institute for Educational Research. Expanded edition, 1980*. Chicago: University of Chicago Press.

Smith, R. M., Schumacker, R. E., & Bush, J. M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66–78.

Tennant, A., & Pallant, J. F. (2012). The root mean square error of approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions, 25*, 1348–1349.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions, 1990*, 84–85.

Chapter 16
Measuring the Effectiveness of Evaluation Processes for Diversified Undergraduate Students with Self-regulated Study—A Case Study in International School of Jinan University

Yufan Liu

Abstract There are more and more universities in China accepting international students from various countries. These students with international background and highly diversified preferences in learning have presented challenges for educators in universities. Using the help of web-based learning platform and being guided in self-regulated study groups, students are now given much more alternatives in terms of learning and communicating. This paper introduces the design of a comprehensive system of teaching evaluating the students of one economic course. The Multi-Facet Rasch Model was applied to measure the effectiveness of such evaluation system.

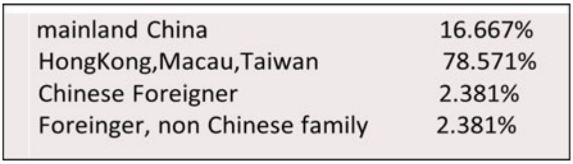
Keywords Web-based learning platform · Self-regulated study · Cooperative study · Multi-facet Rasch model

16.1 Introduction

Since the beginning of the twenty first century, there were more and more universities in China accepting international students from various countries. In 2013 there were 356,499 students from 200 countries or regions to study in China, with 147, 890 (41.48 %) came for degree programs (Education 2013). In Jinan

Y. Liu (✉)
Jinan University, 510632 Guangzhou, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_16



mainland China	16.667%
HongKong,Macau,Taiwan	78.571%
Chinese Foreigner	2.381%
Foreinger, non Chinese family	2.381%

Fig. 16.1 Diversified student body. *Source* Data from survey on 220 students in three different majors

University, Guangzhou, there were 11,532 international students (24.67 % of the total enrolled) from 142 different countries or regions (including Hong Kong, Macau, and Taiwan) in 2013.[1]

There is an exceptional example in the International School of Jinan University, in 2013, near 1700 students from 75 different countries and regions enrolled in eight different non-language undergraduate programs. These students, with international background and highly diversified preferences in learning, have presented challenges for educators in universities. While many universities still apply traditional lecture-exam type of teaching to these international students, the results are not satisfactory. The unsatisfactory results can partly be seen from the poor exam scores received by international students. Very few international students can adapt to the teaching style of tradition formats that consists of only lectures and exams. The feedbacks received from the students also reflect the dissatisfaction of the students.

A survey (Fig. 16.1) conducted by the author in 2010 on international students' preferences regarding study showed that highly diversified student origins also came with:

- Diversified objectives of learning: Some students plan to continue study in graduate programs, while some only intend to complete a bachelor's degree and start working. Some students wish to continue with their own businesses while studying for the degree, mainly to acquire contacts.
- Diversified background of education: Standards of middle school education vary from country to country. Some students have only completed 4 years of middle school while in China it is 6 years. A 2 year difference in education creates obvious gap in knowledge and ability. Those who realize the shortage of their knowledge will need to make up by additional study.
- Diversified knowledge in economics/business: Students have diverse knowledge in economics, which is the subject of research in this paper. In most developed countries, "principles of economics" is taught at high school level. In less developed countries, students have no training at all in this field.
- Diversified skills of language. Since the teaching language is English in the school where this study is conducted, the abilities to read, write, listen, and

[1] Data provided by the Dean's office of International School, Jinan University.

speak in English are crucial. Some students come from English-speaking countries, but some students can only hardly pass preliminary entrance exam.
- Diversified ability in quantitative analysis. Mathematics and statistics pose to be difficult for almost all students in economics. While Chinese students can do very well in math, students from other countries struggle with it all throughout the university study.

Although separated teaching at different levels of abilities was one of the alternatives to improve the result of education, students did not think of separated teaching according to levels of knowledge or skills a good option for major courses, due to the nature of the major they study. For students in business and economics, personal contact is of great importance.

The same survey also revealed some important preferences by the students:

- Lectures by teacher (90.5 %). It was widely thought that students would not like to be constrained in the classroom and listen to the teachers' lectures. But in fact they would like to be lectured. To understand the concepts and theories, students felt that good lectures by teachers could help to make their study more efficient.
- Guidance by teacher (89.4 %). In the process of conducting self-regulated study, students felt the need to be guided by certain rule or recommendations. By giving students necessary rubrics for self-regulated study activities, the teacher could have better control over the process and have outcome better meeting expectation.
- Group discussions (90.5 %). The format of group study was favored by most students. In a class setting with diversified students, they enjoyed the atmosphere brought by peers with different origins, interest, life styles, and accents. Foreign students were found to enjoy contacts with Chinese students for the benefit of possible business opportunities, while Chinese students liked to have contact with foreign students for better adaptability in future careers.
- Using more illustrative tools such as charts and graphics (95.3 %). Due to the fact that students with less mathematical skills will prefer graphical illustrations than pure mathematical expressions, such preference could be adopted in the teaching materials to help with comprehension.

16.2 Developing the Evaluation System

In order to improve the results of education, the author and a team of teachers have conducted several reforms in teaching economic courses to international students. One of the attempts was using the help of web-based learning platform and self-regulated study groups to help build a comprehensive system of learning and evaluation.

Studies on self-regulated study, cooperative study, and web-based study can be found in numerous sources. Chen and Zhou (2009) found positive relationship between time management, short-term planning, long-term planning, general self-efficacy, and the effectiveness of self-regulated study. Studies also found that motivation, ability to solve problems, ability in utilizing computers, and the monitoring and controlling of the study process were the most important factors affecting the results of self-study (Chen and Chen 2010; Li and Xu 2010). Students conduct self-regulated study usually in the forms of cooperative study and computer-aided studies. While many Chinese scholars were studying the factors determining the effectiveness of self-regulated study, many from outside of China have been concerned about the positive impacts of computer utilization, internet connection, and cooperation with peers on effective study (Gokhale 1995; Horsburgh et al. 2001; Moller et al. 2005). The web-based applications and platforms enabled students to communicate extensively and perform cross-discipline cooperations (Hines et al. 1998). Studies in New Zealand universities have found that the students, especially international students would love to perform cooperated studies in self-regulated groups. This form of study benefited students groups that contain diversified backgrounds and motivations (Jeffrey 2009).

Scholars have also found that simply providing web-based platform and letting the students to form self-regulated groups are not enough. There should also be supervision and guidance by teachers to help the students to avoid unnecessary information and unwanted misunderstandings (Caballé et al. 2010; Demetriadis et al. 2008). Not only supervision and guidance, but also face-to-face communications, in-classroom discussions and lectures were found to be essential in making self-regulated study successful (Lytras et al. 2009; Stahl 2006; Wang et al. 2008).

Based on the insights from other scholars' research results, the author designed a course that contains contents and achievements from different activities. The key concepts and difficult contents were lectures by teacher in class. The preview, practice, reading, or review exercises were given in forms of assignments, group projects, online quizzes, etc. A written final exam would be given at the end of semester. The contents in these activities had different levels of difficulty. The easy contents must be mastered, intermediate contents needed comprehension only, while difficult contents would not be tested. Students were given much more alternatives in terms of learning and communicating.

As part of the reform, students were given various types of evaluation: written assignments, online quizzes, group projects, and written exams. Among the above items, written assignments and online quizzes were independent works. Group projects needed to be completed by cooperation, as well as according to the guidance and requirements given by the teacher. Written exam was the final exam at the end of the semester. The results of such complex evaluation system gave students more opportunities to develop skills in comprehension, analysis, cooperation, and communication. Figure 16.2 is an illustration of the interrelated components in this evaluation process, which is centered on classroom lectures. As shown in the figure, teachers have control over the system by monitoring and guiding the students through all these activities. It is also important for the students to be able to connect to teacher and other students through this system.

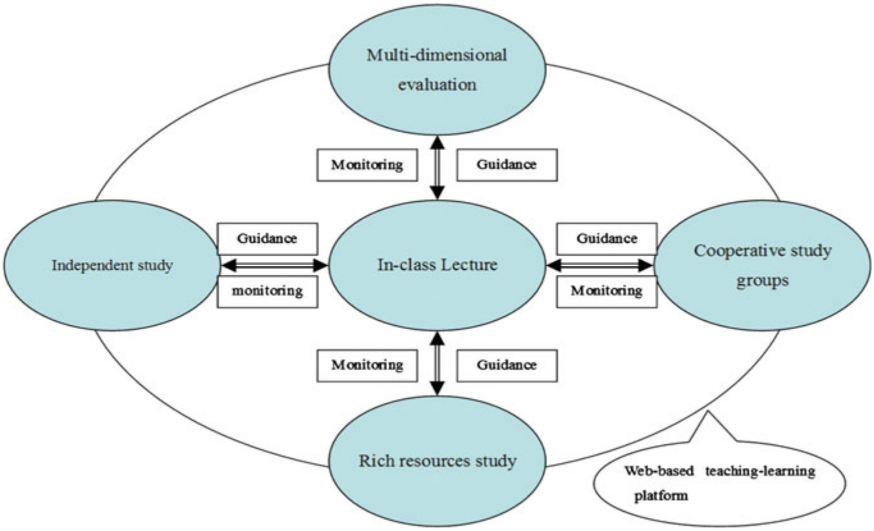


Fig. 16.2 An inter-related system of self-regulated study and evaluation

16.3 Applying the Many-Facets Rasch Model (MFRM) to Measure the Effectiveness of Evaluation

After the students were evaluated by the system that contains various criteria, it is necessary to measure the effectiveness of such system. The many-facets Rasch model has been applied in measuring abilities in language study or performance assessments in various subjects. He and Zhang applied the many-facets Rasch model to measure the reliability of CET (College English Test) in China, using test scores of students from one particular exam (He and Zhang 2008). The analysis of data revealed reliability of such test, by finding statistical significance among different facets: rater severity, task difficulty, rating criteria, and rating scale. Wang and Huang used the MFRM to measure student performance in one classroom activity, with students rating each other's performances. In this study, the reliability and differences of raters and rating scales were analyzed (Wang and Huang 2013). In a different setting, where students were asked to rate teacher's teaching quality, the MFRM was used to find the rater severity, criterion difficulty, and scale significance in the evaluating process (Gao 2012). Sun applied classical test theory and found the insignificance of testing scores with respect to the abilities of students in an English exam. But it was by applying the MFRM Sun was able to identify the reason for such insignificance. The MFRM reveal that the imbalance in the rater severity, difficulty of tasks, and bias iteration between some students and tasks that had caused the insignificance (Sun 2010). The above listed studies have shown the ability of MFRM in revealing the differences in rater severity, task difficulty, and demographic characteristics. This study intends to measure the effectiveness of an evaluation system that contains examinees, testing tasks, and the demographic characteristics. Therefore, an MFRM seemed quite suitable for this purpose.

16.4 Data

This study took the results of 118 students, who studied the course of Public Finance. The first facet was the students enrolled in the course. They were labeled with numbers 1–118. The second facet was the region the students were from. The classification of the regions were: Mainland China (labeled "Mainland"), Hong Kong, Macau, and Taiwan (labeled "Hong Kong"), Chinese origin foreigners (labeled "Chinese Foreign"), other foreigners (labeled "Foreign"). Composition of the students could be seen in Table 16.1.

There were nine components, aiming for a list of abilities to be evaluated (Table 16.2).

Students received marking on each of the nine criteria, in various types of scales: some in percentage points and some in 100 points and some in 5 points. In the data collecting process, scaling of each criterion was adjusted to 0–10 scale. The weights of these criteria were set to be equal.

16.5 Analysis

16.5.1 Disjoint Subsets

There were four disjoint subsets indicated by the initial run of MFRM measuring (Table 16.3a, b). The reason was that students from different regions showed significant differences in abilities of studying. The different regions' education systems were different and students had not been given a uniformed standard of academic abilities.

Table 16.1 Composition of students

Regions of origin	Label of region	Number of students	Percentage of total students (%)
Mainland China	Mainland	34	28.81
Foreigners	Foreign	53	44.92
Foreigners with Chinese origin	Chinese Foreign	14	11.86
Hong Kong/Macau/Taiwan	Hong Kong	17	14.41

Table 16.2 Components of evaluation criteria—the tests

Objectives of tests	Allocation of objectives in various tests	Labels of tests
Understanding of basic concepts and theories	Multiple choice questions—final exam Assignments Online quizzes	MC Assignment Online Quizzes
Application of theories in real life issues	Short essay questions—final exam Presentation—group project Assignments	SE Presentation
Quantitative analysis	Problem solving questions—final exam Online quizzes Assignments	PR
Logical thinking	True/false questions—final exam Online quizzes	T/F
Amount of reading and searching	Essay writing—final exam Presentation—group project	COM
Cooperation with others	Presentation—group project	
Communication with public	Presentation—group project	
Business conduct	Assignments Presentation—group project	
Participation	Attendance records	Attendance

To take into consideration of the obvious difference in the four groups of students, anchoring of the elements in the student and criteria facets seemed to be necessary. The students could not choose where they came from or how to receive their primary and secondary education, and the difference in the origin certainly contributed to the differences in academic performance. These differences could not easily be smoothed out by a short period of study, especially when the students are just sophomores.

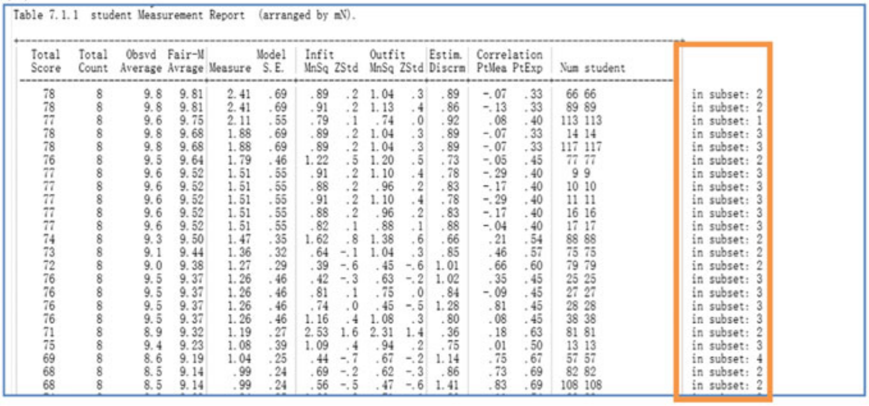
The anchoring process anchored the two facets: students and evaluation criteria. The "Region" facet was not anchored. The rerun of the analysis reported "subset connection OK". But the downside of anchoring is that data might be over-constrained and raw-score difference was greater than 0.5, as shown in the result of this analysis in Table 16.4.

16.5.2 Global Measurement Statistics

The measurable data summary reveals that both mean residual and standardized mean residual were zero. The population standard deviation was 1.04, where all possible elements were included. Chi-square significance probability $p = 0.0000$. Also that 72.14 % of the variance was explained by Rasch measurement, which seemed satisfactory in terms of overall fit at the moment (Tables 16.5, 16.6 and 16.7).

Table 16.3 Disjointed subsets indicated

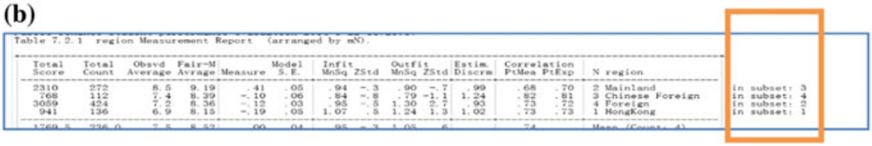
(a)



```
Table 7.1.1  student Measurement Report  (arranged by mN).

+--------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|         Model | Infit      Outfit   |Estim.| Correlation |              
|  Score   Count  Average Avrage|Measure   S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Num student  
|-------------------------------+---------------+---------------------+------+-------------+--------------
|     78       8      9.8   9.81|   2.41    .69 |  .89   .2  1.04   .3|  .89 |  -.07   .33 |  66 66        in subset: 2
|     78       8      9.8   9.81|   2.41    .69 |  .91   .2  1.13   .4|  .86 |  -.13   .33 |  89 89        in subset: 2
|     77       8      9.6   9.75|   2.11    .55 |  .79   .1   .74   .0|  .92 |   .08   .40 | 113 113       in subset: 1
|     78       8      9.8   9.68|   1.88    .69 |  .89   .2  1.04   .3|  .89 |  -.07   .33 |  14 14        in subset: 3
|     78       8      9.8   9.68|   1.88    .69 |  .89   .2  1.04   .3|  .89 |  -.07   .33 | 117 117       in subset: 3
|     76       8      9.5   9.64|   1.79    .46 | 1.22   .5  1.20   .5|  .73 |  -.05   .45 |  77 77        in subset: 2
|     77       8      9.6   9.52|   1.51    .55 |  .91   .2  1.10   .4|  .78 |  -.29   .40 |   9 9         in subset: 3
|     77       8      9.6   9.52|   1.51    .55 |  .88   .2   .96   .2|  .83 |  -.17   .40 |  10 10        in subset: 3
|     77       8      9.6   9.52|   1.51    .55 |  .91   .2  1.10   .4|  .78 |  -.29   .40 |  11 11        in subset: 3
|     77       8      9.6   9.52|   1.51    .55 |  .88   .2   .96   .2|  .83 |  -.17   .40 |  16 16        in subset: 3
|     77       8      9.6   9.52|   1.51    .55 |  .82   .1   .88   .1|  .88 |  -.04   .40 |  17 17        in subset: 3
|     74       8      9.3   9.50|   1.47    .35 | 1.62   .8  1.38   .6|  .66 |   .21   .54 |  88 88        in subset: 2
|     73       8      9.1   9.44|   1.36    .32 |  .64  -.1  1.04   .3|  .85 |   .46   .57 |  75 75        in subset: 2
|     72       8      9.0   9.38|   1.27    .29 |  .39  -.6   .45  -.6| 1.01 |   .66   .60 |  79 79        in subset: 2
|     76       8      9.5   9.37|   1.26    .46 |  .42  -.3   .63  -.2| 1.02 |   .35   .45 |  25 25        in subset: 3
|     76       8      9.5   9.37|   1.26    .46 |  .81   .1   .75   .0|  .84 |  -.09   .45 |  27 27        in subset: 3
|     76       8      9.5   9.37|   1.26    .46 |  .74   .0   .45  -.5| 1.28 |   .81   .45 |  28 28        in subset: 3
|     76       8      9.5   9.37|   1.26    .46 | 1.16   .4  1.08   .3|  .80 |   .08   .45 |  38 38        in subset: 3
|     71       8      8.9   9.32|   1.19    .27 | 2.53  1.6  2.31  1.4|  .36 |   .18   .63 |  81 81        in subset: 2
|     75       8      9.4   9.23|   1.08    .39 | 1.09   .4   .94   .2|  .75 |   .01   .50 |  13 13        in subset: 3
|     69       8      8.6   9.19|   1.04    .25 |  .44  -.7   .67  -.2| 1.14 |   .75   .67 |  57 57        in subset: 4
|     68       8      8.5   9.14|    .99    .24 |  .69  -.2   .62  -.3|  .86 |   .73   .69 |  82 82        in subset: 2
|     68       8      8.5   9.14|    .99    .24 |  .56  -.5   .47  -.6| 1.41 |   .83   .69 | 108 108       in subset: 2
```

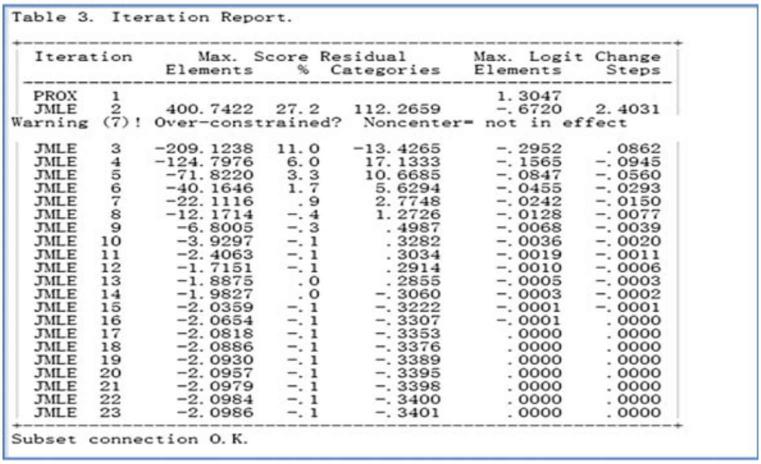
(b)



```
Table 7.2.1  region Measurement Report  (arranged by mN).

+--------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|         Model | Infit      Outfit   |Estim.| Correlation |                     
|  Score   Count  Average Avrage|Measure   S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N region            
|-------------------------------+---------------+---------------------+------+-------------+---------------------
|   2310     272      8.5   9.19|    .41    .05 |  .94  -.3   .90  -.7|  .99 |   .68   .70 | 2 Mainland           in subset: 3
|    768     112      7.4   8.39|   -.10    .06 |  .84  -.8   .79 -1.1| 1.24 |   .82   .81 | 3 Chinese Foreign    in subset: 4
|   3059     424      7.2   8.36|   -.12    .03 |  .95  -.5  1.30  2.7|  .93 |   .73   .72 | 4 Foreign            in subset: 2
|    941     136      6.9   8.15|   -.19    .05 | 1.07   .5  1.24  1.3| 1.02 |   .73   .73 | 1 HongKong           in subset: 1
|-------------------------------+---------------+---------------------+------+-------------+---------------------
| 1769.5   336.0      7.5   8.52|    .00    .04 |  .95  -.3  1.05   .6|      |   .74       | Mean (Count: 4)
```

Table 16.4 Iteration report with anchoring

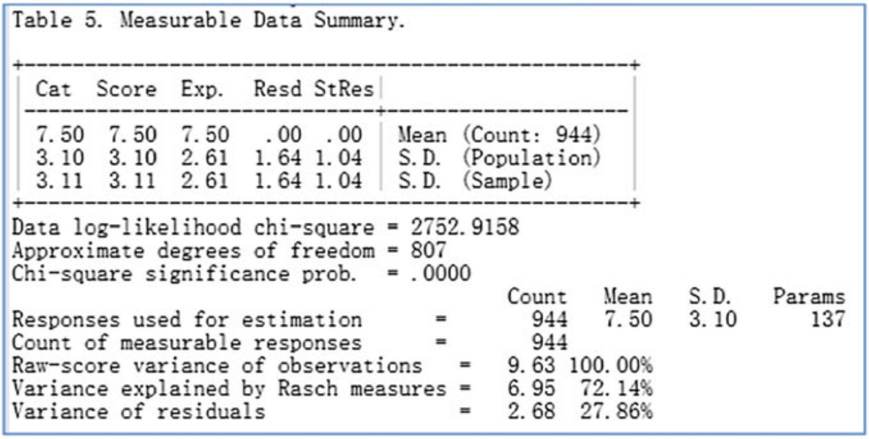


```
Table 3.  Iteration Report.

+----------------------------------------------------------------------------+
| Iteration        Max.  Score Residual          Max.  Logit Change          |
|               Elements    %   Categories    Elements       Steps          |
|----------------------------------------------------------------------------|
| PROX     1                                      1.3047                      |
| JMLE     2    400.7422   27.2    112.2659      -.6720       2.4031          |
Warning (7)! Over-constrained?  Noncenter= not in effect

| JMLE     3   -209.1238   11.0    -13.4265      -.2952        .0862          |
| JMLE     4   -124.7976    6.0     17.1333      -.1565       -.0945          |
| JMLE     5    -71.8220    3.3     10.6685      -.0847       -.0560          |
| JMLE     6    -40.1646    1.7      5.6294      -.0455       -.0293          |
| JMLE     7    -22.1116     .9      2.7748      -.0242       -.0150          |
| JMLE     8    -12.1714    -.4      1.2726      -.0128       -.0077          |
| JMLE     9     -6.8005    -.3       .4987      -.0068       -.0039          |
| JMLE    10     -3.9297    -.1       .3282      -.0036       -.0020          |
| JMLE    11     -2.4063    -.1       .3034      -.0019       -.0011          |
| JMLE    12     -1.7151    -.1       .2914      -.0010       -.0006          |
| JMLE    13     -1.8875     .0       .2855      -.0005       -.0003          |
| JMLE    14     -1.9827     .0      -.3060      -.0003       -.0002          |
| JMLE    15     -2.0359    -.1      -.3222      -.0001       -.0001          |
| JMLE    16     -2.0654    -.1      -.3307      -.0001        .0000          |
| JMLE    17     -2.0818    -.1      -.3353       .0000        .0000          |
| JMLE    18     -2.0886    -.1      -.3376       .0000        .0000          |
| JMLE    19     -2.0930    -.1      -.3389       .0000        .0000          |
| JMLE    20     -2.0957    -.1      -.3395       .0000        .0000          |
| JMLE    21     -2.0979    -.1      -.3398       .0000        .0000          |
| JMLE    22     -2.0984    -.1      -.3400       .0000        .0000          |
| JMLE    23     -2.0986    -.1      -.3401       .0000        .0000          |
+----------------------------------------------------------------------------+
Subset connection O.K.
```

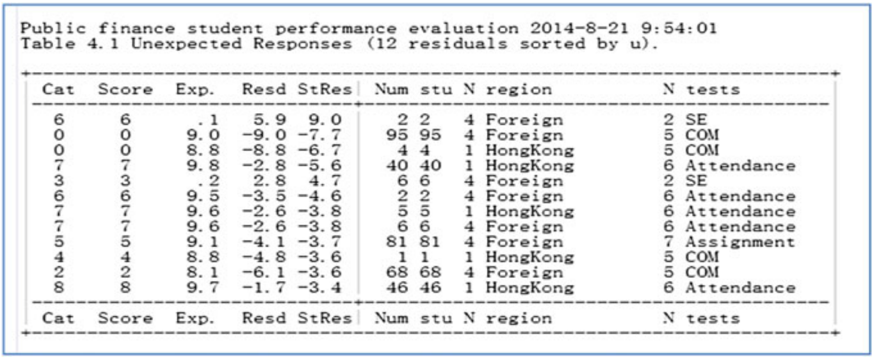
Table 16.5 Measurable data summary



```
Table 5. Measurable Data Summary.

+-----------------------------------------------------+
| Cat   Score   Exp.   Resd StRes|                     |
--------------------------------+--------------------
 7.50  7.50  7.50    .00   .00 | Mean (Count: 944)
 3.10  3.10  2.61   1.64  1.04 | S.D. (Population)
 3.11  3.11  2.61   1.64  1.04 | S.D. (Sample)
+-----------------------------------------------------+
Data log-likelihood chi-square = 2752.9158
Approximate degrees of freedom = 807
Chi-square significance prob.  = .0000
                                      Count   Mean   S.D.   Params
Responses used for estimation    =     944    7.50   3.10      137
Count of measurable responses    =     944
Raw-score variance of observations  =  9.63 100.00%
Variance explained by Rasch measures =  6.95  72.14%
Variance of residuals               =  2.68  27.86%
```

Table 16.6 Unexpected responses with “attendance” in data set, $\mu > 3$



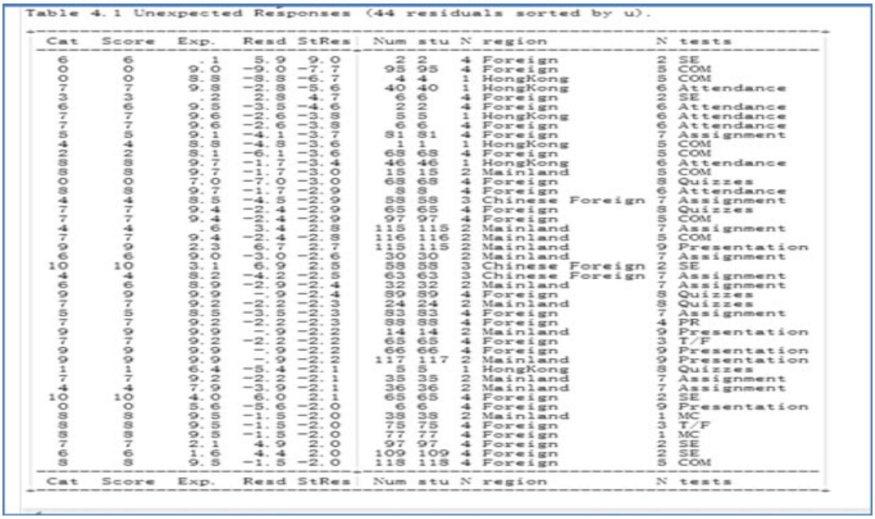
```
Public finance student performance evaluation 2014-8-21 9:54:01
Table 4.1 Unexpected Responses (12 residuals sorted by u).

+-----------------------------------------------------------------------+
| Cat   Score   Exp.   Resd StRes| Num stu N region           N tests
-----------------------------------+------------------------------------
   6      6      .1    5.9   9.0 |   2  2   4 Foreign         2 SE
   0      0     9.0   -9.0  -7.7 |  95 95   4 Foreign         5 COM
   0      0     8.8   -8.8  -6.7 |   4  4   1 HongKong        5 COM
   7      7     9.8   -2.8  -5.6 |  40 40   1 HongKong        6 Attendance
   3      3      .2    2.8   4.7 |   6  6   4 Foreign         2 SE
   6      6     9.5   -3.5  -4.6 |   2  2   4 Foreign         6 Attendance
   7      7     9.6   -2.6  -3.8 |   5  5   1 HongKong        6 Attendance
   7      7     9.6   -2.6  -3.8 |   6  6   4 Foreign         6 Attendance
   5      5     9.1   -4.1  -3.7 |  81 81   4 Foreign         7 Assignment
   4      4     8.8   -4.8  -3.6 |   1  1   1 HongKong        5 COM
   2      2     8.1   -6.1  -3.6 |  68 68   4 Foreign         5 COM
   8      8     9.7   -1.7  -3.4 |  46 46   1 HongKong        6 Attendance
-----------------------------------+------------------------------------
| Cat   Score   Exp.   Resd StRes| Num stu N region           N tests
+-----------------------------------------------------------------------+
```

16.5.3 Unexpected Responses and the Fitting of Model

When all nine criteria were included in the data set, 12 unexpected responses $\mu > 3$, which is 12/1062 = 1.1 %. 66 unexpected responses by $\mu > 2$, which is 66/1062 = 6.2 %. The data set contains one evaluation criterion that is not given by judgment or testing: attendance. The rating of this criterion was given, rather than by evaluation, by records of student attendance, which is a mandatory rule set by the university. It seems more reasonable to remove this criterion, considering how it is determined. The modified unexpected response reports are shown in Tables 16.8 and 16.9: 7 unexpected responses with $\mu > 2$, which is 7/944 = 0.74 %; 39

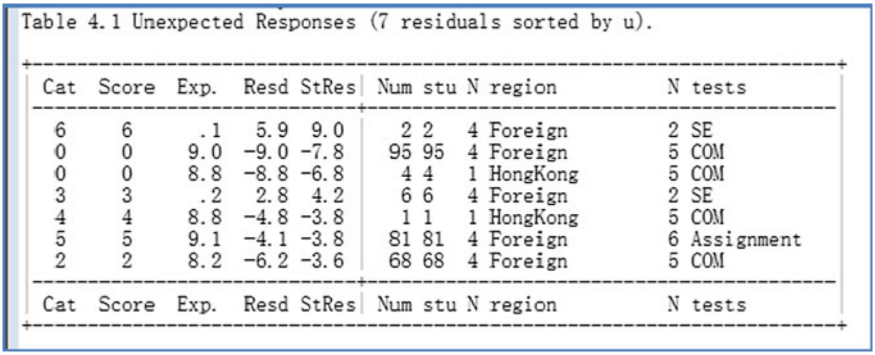
Table 16.7 Unexpected responses with “attendance” in data set, $\mu > 2$



```
Table 4.1 Unexpected Responses (44 residuals sorted by u).

+-------------------------------------------------------------------------------
| Cat  Score  Exp.   Resd StRes| Num stu N region              N tests
|------------------------------+-------------------------------------------------
|  6     6      .1    5. 9  9. 0|   2   2  4 Foreign             2 SE
|  0     0    9. 0  -9. 0 -7. 7|  95  95  4 Foreign             5 COM
|  0     0    8. 8  -8. 8 -6. 7|   4   4  1 HongKong            5 COM
|  7     7    9. 8  -2. 8 -5. 6|  40  40  1 HongKong            6 Attendance
|  3     3      .2    2. 8  4. 7|   6   6  4 Foreign             2 SE
|  6     6    9. 5  -3. 5 -4. 6|   2   2  4 Foreign             6 Attendance
|  7     7    9. 6  -2. 6 -3. 8|   5   5  1 HongKong            6 Attendance
|  7     7    9. 6  -2. 6 -3. 8|   6   6  4 Foreign             6 Attendance
|  5     5    9. 1  -4. 1 -3. 7|  81  81  4 Foreign             7 Assignment
|  4     4    8. 8  -4. 8 -3. 6|   1   1  1 HongKong            5 COM
|  2     2    8. 1  -6. 1 -3. 6|  68  68  4 Foreign             5 COM
|  8     8    9. 7  -1. 7 -3. 4|  46  46  1 HongKong            6 Attendance
|  8     8    9. 7  -1. 7 -3. 0|  15  15  2 Mainland            5 COM
|  0     0    7. 0  -7. 0 -3. 0|  68  68  4 Foreign             8 Quizzes
|  8     8    9. 7  -1. 7 -2. 9|   8   8  4 Foreign             6 Attendance
|  4     4    8. 5  -4. 5 -2. 9|  58  58  3 Chinese Foreign     7 Assignment
|  7     7    9. 4  -2. 4 -2. 9|  65  65  4 Foreign             8 Quizzes
|  7     7    9. 4  -2. 4 -2. 9|  97  97  4 Foreign             5 COM
|  4     4      .6    3. 4  2. 8| 115 115  2 Mainland            7 Assignment
|  7     7    9. 4  -2. 4 -2. 8| 116 116  2 Mainland            5 COM
|  9     9    2. 3    6. 7  2. 7| 115 115  2 Mainland            9 Presentation
|  6     6    9. 0  -3. 0 -2. 6|  30  30  2 Mainland            7 Assignment
| 10    10    3. 1    6. 9  2. 5|  58  58  3 Chinese Foreign     2 SE
|  4     4    8. 2  -4. 2 -2. 5|  63  63  3 Chinese Foreign     7 Assignment
|  6     6    8. 9  -2. 9 -2. 4|  32  32  2 Mainland            7 Assignment
|  9     9    9. 9   -. 9 -2. 4|  89  89  4 Foreign             8 Quizzes
|  7     7    9. 2  -2. 2 -2. 3|  24  24  2 Mainland            8 Quizzes
|  5     5    8. 5  -3. 5 -2. 3|  83  83  4 Foreign             7 Assignment
|  7     7    9. 2  -2. 2 -2. 3|  88  88  4 Foreign             4 PR
|  9     9    9. 9   -. 9 -2. 2|  14  14  2 Mainland            9 Presentation
|  7     7    9. 2  -2. 2 -2. 2|  65  65  4 Foreign             3 T/F
|  9     9    9. 9   -. 9 -2. 2|  66  66  4 Foreign             9 Presentation
|  9     9    9. 9   -. 9 -2. 2| 117 117  2 Mainland            9 Presentation
|  1     1    6. 4  -5. 4 -2. 1|   5   5  1 HongKong            8 Quizzes
|  7     7    9. 2  -2. 2 -2. 1|  35  35  2 Mainland            7 Assignment
|  4     4    7. 9  -3. 9 -2. 1|  36  36  2 Mainland            7 Assignment
| 10    10    4. 0    6. 0  2. 1|  65  65  4 Foreign             2 SE
|  0     0    5. 6  -5. 6 -2. 0|   6   6  4 Foreign             9 Presentation
|  8     8    9. 5  -1. 5 -2. 0|  38  38  2 Mainland            1 MC
|  8     8    9. 5  -1. 5 -2. 0|  75  75  4 Foreign             3 T/F
|  8     8    9. 5  -1. 5 -2. 0|  77  77  4 Foreign             1 MC
|  7     7    2. 1    4. 9  2. 0|  97  97  4 Foreign             2 SE
|  6     6    1. 6    4. 4  2. 0| 109 109  4 Foreign             2 SE
|  8     8    9. 5  -1. 5 -2. 0| 118 118  4 Foreign             5 COM
|------------------------------+-------------------------------------------------
| Cat  Score  Exp.   Resd StRes| Num stu N region              N tests
+-------------------------------------------------------------------------------
```

Table 16.8 Unexpected responses without “attendance” in data set, $\mu > 3$

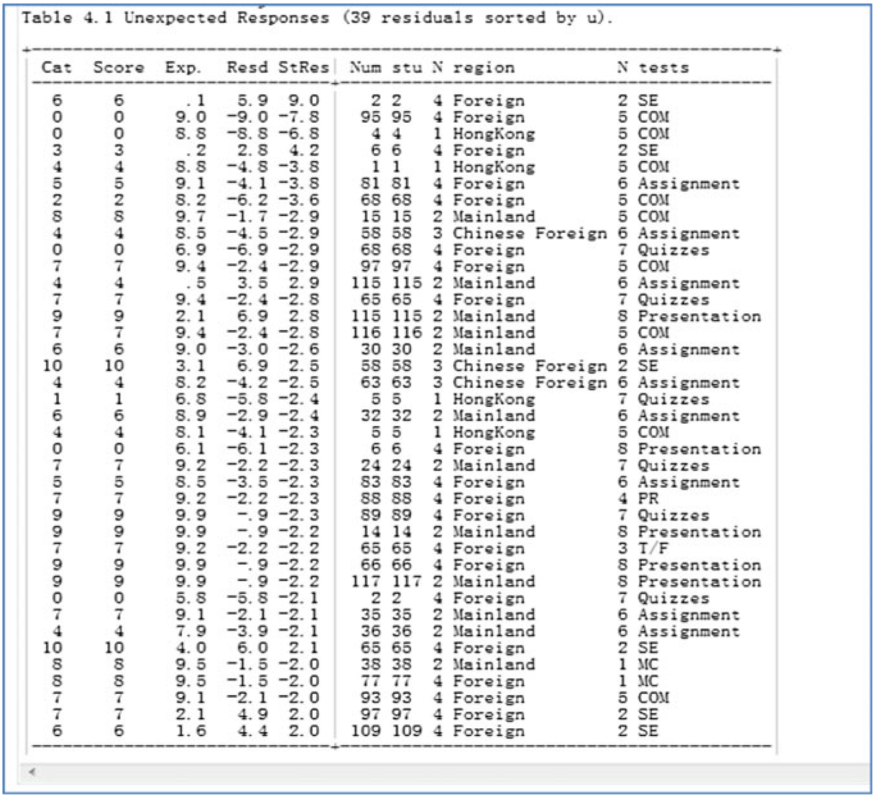


```
Table 4.1 Unexpected Responses (7 residuals sorted by u).

+--------------------------------------------------------------------------+
| Cat  Score  Exp.   Resd StRes| Num stu N region           N tests        |
|------------------------------+-------------------------------------------|
|  6     6      .1    5. 9  9. 0|   2 2   4 Foreign          2 SE           |
|  0     0    9. 0  -9. 0 -7. 8|  95 95  4 Foreign          5 COM          |
|  0     0    8. 8  -8. 8 -6. 8|   4 4   1 HongKong         5 COM          |
|  3     3      .2    2. 8  4. 2|   6 6   4 Foreign          2 SE           |
|  4     4    8. 8  -4. 8 -3. 8|   1 1   1 HongKong         5 COM          |
|  5     5    9. 1  -4. 1 -3. 8|  81 81  4 Foreign          6 Assignment   |
|  2     2    8. 2  -6. 2 -3. 6|  68 68  4 Foreign          5 COM          |
|------------------------------+-------------------------------------------|
| Cat  Score  Exp.   Resd StRes| Num stu N region           N tests        |
+--------------------------------------------------------------------------+
```

unexpected responses with $\mu > 3$, which is 39/944 = 4.13 %. The specific records of unexpected responses were all accounted for with reasons such as absent for lectures, absent for final exam, or failed to submit assignments, etc. For example, student number 2 had failed to submit assignments, online quizzes, and did not take part in presentation, therefore failed the course. The high score received in the short essay question was pure luck. Student number 95 had achieved high marks on assignments and online quizzes, but failed short essays, problem solving, and comprehensive problem sections. Apparently this student did not do the assignments or quizzes honestly. Students 4, 6, and 1 failed the final exam due to poor overall performance

Table 16.9 Unexpected responses without “attendance” in data set, $\mu > 2$



```
Table 4.1 Unexpected Responses (39 residuals sorted by u).

+--------------------------------------------------------------------------------+
| Cat  Score  Exp.   Resd StRes| Num stu N region           N tests               |
|-----------------------------+-------------------------------------------------|
   6     6      .1    5.9   9.0    2 2    4 Foreign           2 SE
   0     0     9.0   -9.0  -7.8   95 95   4 Foreign           5 COM
   0     0     8.8   -8.8  -6.8    4 4    1 HongKong          5 COM
   3     3      .2    2.8   4.2    6 6    4 Foreign           2 SE
   4     4     8.8   -4.8  -3.8    1 1    1 HongKong          5 COM
   5     5     9.1   -4.1  -3.8   81 81   4 Foreign           6 Assignment
   2     2     8.2   -6.2  -3.6   68 68   4 Foreign           5 COM
   8     8     9.7   -1.7  -2.9   15 15   2 Mainland          5 COM
   4     4     8.5   -4.5  -2.9   58 58   3 Chinese Foreign 6 Assignment
   0     0     6.9   -6.9  -2.9   68 68   4 Foreign           7 Quizzes
   7     7     9.4   -2.4  -2.9   97 97   4 Foreign           5 COM
   4     4      .5    3.5   2.9  115 115 2 Mainland          6 Assignment
   7     7     9.4   -2.4  -2.8   65 65   4 Foreign           7 Quizzes
   9     9     2.1    6.9   2.8  115 115 2 Mainland          8 Presentation
   7     7     9.4   -2.4  -2.8  116 116 2 Mainland          5 COM
   6     6     9.0   -3.0  -2.6   30 30   2 Mainland          6 Assignment
  10    10     3.1    6.9   2.5   58 58   3 Chinese Foreign 2 SE
   4     4     8.2   -4.2  -2.5   63 63   3 Chinese Foreign 6 Assignment
   1     1     6.8   -5.8  -2.4    5 5    1 HongKong          7 Quizzes
   6     6     8.9   -2.9  -2.4   32 32   2 Mainland          6 Assignment
   4     4     8.1   -4.1  -2.3    5 5    1 HongKong          5 COM
   0     0     6.1   -6.1  -2.3    6 6    4 Foreign           8 Presentation
   7     7     9.2   -2.2  -2.3   24 24   2 Mainland          7 Quizzes
   5     5     8.5   -3.5  -2.3   83 83   4 Foreign           6 Assignment
   7     7     9.2   -2.2  -2.3   88 88   4 Foreign           4 PR
   9     9     9.9    -.9  -2.3   89 89   4 Foreign           7 Quizzes
   9     9     9.9    -.9  -2.2   14 14   2 Mainland          8 Presentation
   7     7     9.2   -2.2  -2.2   65 65   4 Foreign           3 T/F
   9     9     9.9    -.9  -2.2   66 66   4 Foreign           8 Presentation
   9     9     9.9    -.9  -2.2  117 117 2 Mainland          8 Presentation
   0     0     5.8   -5.8  -2.1    2 2    4 Foreign           7 Quizzes
   7     7     9.1   -2.1  -2.1   35 35   2 Mainland          6 Assignment
   4     4     7.9   -3.9  -2.1   36 36   2 Mainland          6 Assignment
  10    10     4.0    6.0   2.1   65 65   4 Foreign           2 SE
   8     8     9.5   -1.5  -2.0   38 38   2 Mainland          1 MC
   8     8     9.5   -1.5  -2.0   77 77   4 Foreign           1 MC
   7     7     9.1   -2.1  -2.0   93 93   4 Foreign           5 COM
   7     7     2.1    4.9   2.0   97 97   4 Foreign           2 SE
   6     6     1.6    4.4   2.0  109 109 4 Foreign           2 SE
|-----------------------------+-------------------------------------------------|
```

in short essays, comprehensive problem, and problem solving. Student 81 passed the course and had performed well in all aspects except for assignments, due to misunderstanding of the deadline. Student 68 failed the course in all criteria, and was not surprised to get only 2 out of 10 in the comprehensive problem.

16.5.4 Student Performances

In the student performance report, 25 students (21.2 % of population) had Infit mean square lower than 0.5, which indicated insignificant level of differentiation. 12 students (10.2 % of population) had Infit mean square higher than 2, which indicated too much differentiation. When outfit mean square was examined, only 18 students (15.3 % of population) had Outfit mean square lower than 0.5 and still 12 students had Outfit mean square higher than 2. The above result showed that in total 25–30 % of the students tested had inadequate levels of differentiation. This percentage was considered satisfactory, comparing with other scholars' studies (Sun 2010). The range of 0.5–2 was suggested by M. Linacre.[2] The contribution for such level of variation might be sufficient number of test items, adequate grading practice of teachers, or just difference in student sources. In order to investigate this issue further, the report of Student Measurement (Tables 16.10 and 16.11) and vertical ruler (Tables 16.12 and 16.13) should be looked into.

Separation ratio was 1.97, and separation reliability ratio was 0.8, indicating 80 percent of the observed variation was properly provided by the test items, rather than due to errors. These two ratios together indicated relatively good separation results given by the evaluation system.

The mean measurement of student performances was 0.64 logit, with a standard deviation of 0.68. Root mean square was 0.31. Students with mean level of ability measure (Table 16.11) are corresponding to the average scale of about 7.5 out of 10. This could be an indication that the average students have somewhat good chance of passing the course, since a log-odd of 0.64 means the ratio of probability of success versus the probability of failure is about 1.89.[3]

In the same table, the facet of region showed that students from mainland China had far better performance than students from other regions. This is not surprising since the Mainland Chinese students were selected out of fierce competition and tough examination. They were the top-achieving students in their high schools. Among students from the other three regions, students who were foreigners but with Chinese origin had better performance than students who are foreigners with no Chinese origin. Students from Hong Kong, Macau, or Taiwan were ranked at the bottom. This outcome is somewhat unexpected, since the quality of Hong Kong, Macau, or Taiwan's education systems were NOT considered poor by many standards. One possibility was that the foreign students who came from various countries had better than average ability in their countries. Overseas Chinese families have always had traditions in paying much attention to education, and their children could be doing quite well in their schools. As for the foreign students, who came to study in China out of their own choices, they could also be those who really were interested in studying and did well. On the contrary, students from Hong Kong, Macau, or Taiwan came to mainland China for not being able to get into local universities or not being able to afford higher education in developed countries such as the United States or U.K. The levels of abilities or motivation were lower than those of the other groups.

[2] M. Linacre's tutorial for Many-facet Rasch Measurements: Facets (1/2012 version).

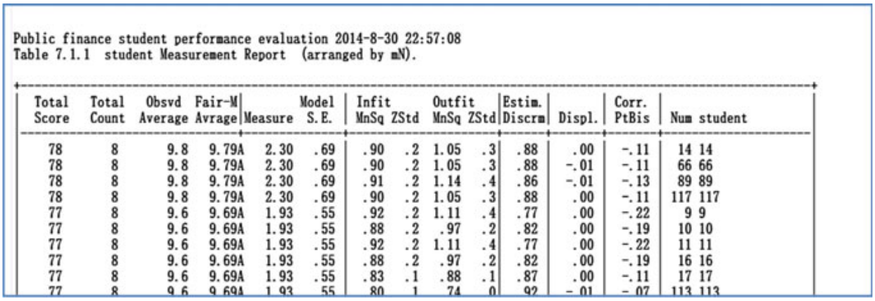
[3] Assuming the chances of success in the course to be P_{ni}, then the log-odd of 0.64 unit would mean for the average-ability student, the chance of passing the course is about 0.65.

$$\ln(P_{\mathrm{ni}}/1-P_{\mathrm{ni}}) = 0.64$$

$$e^{0.64} = P_{\mathrm{ni}}/1-P_{\mathrm{ni}}$$

$$P_{\mathrm{ni}} = e^{0.64}/(1+e^{0.64}) = 0.6546$$

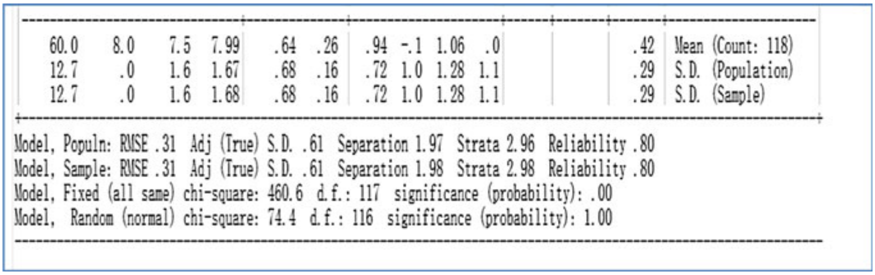
Table 16.10 Header of student measurement report



```
Public finance student performance evaluation 2014-8-30 22:57:08
Table 7.1.1  student Measurement Report  (arranged by mN).

+-------------------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair-M|          Model | Infit      Outfit    |Estim.|        | Corr. |              |
| Score   Count  Average Avrage|Measure   S.E.  | MnSq ZStd  MnSq ZStd|Discrm| Displ. | PtBis | Num student  |
|-------------------------------+----------------+---------------------+------+--------+-------+--------------|
|   78      8       9.8  9.79A   2.30    .69   |  .90   .2  1.05   .3|  .88 |   .00  |  -.11 |  14 14       |
|   78      8       9.8  9.79A   2.30    .69   |  .90   .2  1.05   .3|  .88 |  -.01  |  -.11 |  66 66       |
|   78      8       9.8  9.79A   2.30    .69   |  .91   .2  1.14   .4|  .86 |  -.01  |  -.13 |  89 89       |
|   78      8       9.8  9.79A   2.30    .69   |  .90   .2  1.05   .3|  .88 |   .00  |  -.11 | 117 117      |
|   77      8       9.6  9.69A   1.93    .55   |  .92   .2  1.11   .4|  .77 |   .00  |  -.22 |   9 9        |
|   77      8       9.6  9.69A   1.93    .55   |  .88   .2   .97   .2|  .82 |   .00  |  -.19 |  10 10       |
|   77      8       9.6  9.69A   1.93    .55   |  .92   .2  1.11   .4|  .77 |   .00  |  -.22 |  11 11       |
|   77      8       9.6  9.69A   1.93    .55   |  .88   .2   .97   .2|  .82 |   .00  |  -.19 |  16 16       |
|   77      8       9.6  9.69A   1.93    .55   |  .83   .1   .88   .1|  .87 |   .00  |  -.11 |  17 17       |
|   77      8       9.6  9.69A   1.93    .55   |  .80   .1   .74   .0|  .92 |  -.01  |  -.07 | 113 113      |
```

Table 16.11 Summary of student measurement report



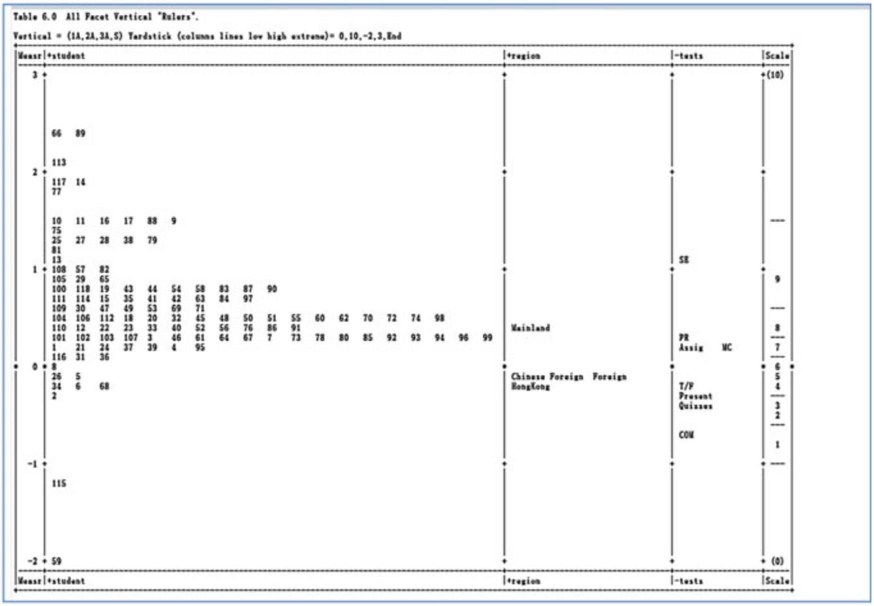
```
|-------------------------------+----------------+---------------------+------+--------+-------+--------------------
|   60.0    8.0     7.5   7.99 |   .64    .26   |  .94  -.1  1.06   .0|      |        |   .42 | Mean (Count: 118)
|   12.7     .0     1.6   1.67 |   .68    .16   |  .72  1.0  1.28  1.1|      |        |   .29 | S.D. (Population)
|   12.7     .0     1.6   1.68 |   .68    .16   |  .72  1.0  1.28  1.1|      |        |   .29 | S.D. (Sample)
+------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .31  Adj (True) S.D. .61  Separation 1.97  Strata 2.96  Reliability .80
Model, Sample: RMSE .31  Adj (True) S.D. .61  Separation 1.98  Strata 2.98  Reliability .80
Model, Fixed (all same) chi-square: 460.6  d.f.: 117  significance (probability): .00
Model,  Random (normal) chi-square: 74.4  d.f.: 116  significance (probability): 1.00
--------------------------------------------------------------------------------------------------------------------
```

The students ranking were modified by anchoring of the region facet. The difference between Tables 16.12 and 16.13 can be observed and summarized as:

- By anchoring the region facet, the estimation process was improved to allow for a smoother iteration. However, the differences of student abilities were still present in both tables.
- The differences between students are slightly more dramatic without anchoring. Students with top rating of measure versus students with bottom rating of measure had a difference of 2.41 − (−2.25) = 4.66 logits, while in the anchored measurements, the distance between top and bottom rating was 2.30 − (−2.36) = 4.56 logits.

In both tables, the top-rated students were mainland students (numbered 117, 14) and foreign students (numbered 66 and 89). Almost all of the mainland students (numbers 9–39, 115–117) were ranked above average except for student number 115, who failed the course due to completely lack of effort. Students number 36, 26, and 34, who were below average, yet above zero, measured at 0.52, 0.29, and 0.26 logits, respectively. Foreign students with Chinese origin (numbered 51–64) had an overall satisfactory performance, which concentrated in the range right above

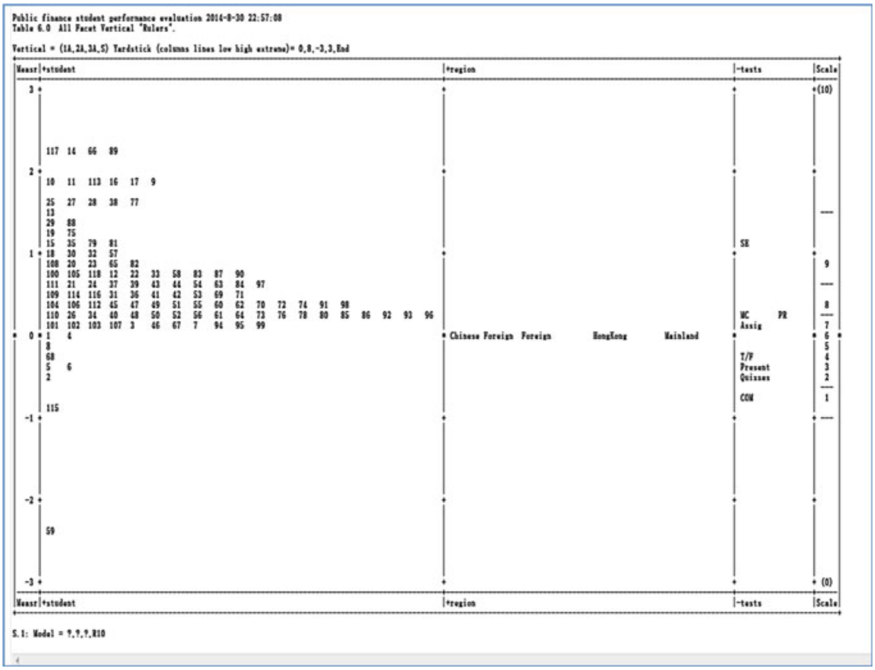
Table 16.12 The vertical rulers of student, region, and test item facets without anchoring



```
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,3A,S) Yardstick (columns lines low high extreme)= 0,10,-2,3,End
+----------------------------------------------------------------------------------------------------------------------------------+
|Measr|+student                                                                       |+region                    |-tests        |Scale|
|-----+------------------------------------------------------------------------------+---------------------------+--------------+-----|
|   3 +                                                                              +                           +              +(10) |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     | 66  89                                                                       |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     | 113                                                                          |                           |              |     |
|   2 +                                                                              +                           +              +     |
|     | 117 14                                                                       |                           |              |     |
|     | 77                                                                           |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     | 10  11  16  17  88  9                                                        |                           |              | --- |
|     | 75                                                                           |                           |              |     |
|     | 25  27  28  38  79                                                           |                           |              |     |
|     | 81                                                                           |                           |              |     |
|     | 13                                                                           |                           | SE           |     |
|   1 + 108 57  82                                                                   +                           +              +     |
|     | 105 29  65                                                                   |                           |              |  9  |
|     | 100 118 19  43  44  54  58  83  87  90                                       |                           |              |     |
|     | 111 114 15  35  41  42  63  84  97                                           |                           |              |     |
|     | 109 30  47  49  53  69  71                                                   |                           |              | --- |
|     | 104 106 112 18  20  32  45  48  50  51  55  60  62  70  72  74  98           |                           |              |     |
|     | 110 12  22  23  33  40  52  56  76  86  91                                   | Mainland                  |              |  8  |
|     | 101 102 103 107 3   46  61  64  67  7   73  78  80  85  92  93  94  96  99   |                           | PR           | --- |
|     | 1   21  24  37  39  4   95                                                   |                           | Assig   MC   |  7  |
|     | 116 31  36                                                                   |                           |              | --- |
*   0 * 8                                                                            *                           *              *  6  *
|     | 26  5                                                                        | Chinese Foreign  Foreign  |              |  5  |
|     | 34  6   68                                                                   | HongKong                  | T/F          |  4  |
|     | 2                                                                            |                           | Present      | --- |
|     |                                                                              |                           | Quizzes      |  3  |
|     |                                                                              |                           |              |  2  |
|     |                                                                              |                           |              | --- |
|     |                                                                              |                           | COM          |     |
|     |                                                                              |                           |              |  1  |
|     |                                                                              |                           |              |     |
|  -1 +                                                                              +                           +              + --- |
|     |                                                                              |                           |              |     |
|     | 115                                                                          |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|     |                                                                              |                           |              |     |
|  -2 + 59                                                                           +                           +              + (0) |
|-----+------------------------------------------------------------------------------+---------------------------+--------------+-----|
|Measr|+student                                                                       |+region                    |-tests        |Scale|
+----------------------------------------------------------------------------------------------------------------------------------+
```

Table 16.13 The vertical rulers of student, region, and test facets, with anchoring



```
Public finance student performance evaluation 2014-8-30 22:57:08
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,3A,S) Yardstick (columns lines low high extreme)= 0,8,-3,3,End
+---------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+student                                                                       |+region                                                    |-tests          |Scale|
|-----+------------------------------------------------------------------------------+-----------------------------------------------------------+----------------+-----|
|   3 +                                                                              +                                                           +                +(10) |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     | 117 14  66  89                                                               |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|   2 +                                                                              +                                                           +                +     |
|     | 10  11  113 16  17  9                                                        |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     | 25  27  28  38  77                                                           |                                                           |                |     |
|     | 13                                                                           |                                                           |                | --- |
|     | 29  88                                                                       |                                                           |                |     |
|     | 19  75                                                                       |                                                           |                |     |
|     | 15  35  79  81                                                               |                                                           | SE             |     |
|   1 + 18  30  32  57                                                               +                                                           +                +     |
|     | 108 20  23  65  82                                                           |                                                           |                |  9  |
|     | 100 105 118 12  22  33  58  83  87  90                                       |                                                           |                |     |
|     | 111 21  24  37  39  43  44  54  63  84  97                                   |                                                           |                | --- |
|     | 109 114 116 31  36  41  42  53  69  71                                       |                                                           |                |     |
|     | 104 106 112 45  47  49  51  55  60  62  70  72  74  91  98                   |                                                           |                |  8  |
|     | 110 26  34  40  48  50  52  56  61  64  73  76  78  80  85  86  92  93  96   |                                                           | MC      PR     | --- |
|     | 101 102 103 107 3   46  67  7   94  95  99                                   |                                                           | Assig          |  7  |
*   0 * 1   4                                                                        * Chinese Foreign  Foreign          HongKong         Mainland  *                *  6  *
|     | 8                                                                            |                                                           |                |  5  |
|     | 68                                                                           |                                                           | T/F            |  4  |
|     | 5   6                                                                        |                                                           | Present        |  3  |
|     | 2                                                                            |                                                           | Quizzes        |  2  |
|     |                                                                              |                                                           |                | --- |
|     |                                                                              |                                                           | COM            |  1  |
|     | 115                                                                          |                                                           |                |     |
|  -1 +                                                                              +                                                           +                + --- |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|  -2 +                                                                              +                                                           +                +     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     | 59                                                                           |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|     |                                                                              |                                                           |                |     |
|  -3 +                                                                              +                                                           +                + (0) |
|-----+------------------------------------------------------------------------------+-----------------------------------------------------------+----------------+-----|
|Measr|+student                                                                       |+region                                                    |-tests          |Scale|
+---------------------------------------------------------------------------------------------------------------------------------------------------------------------+

S.1: Model = ?,?,?,R10
```

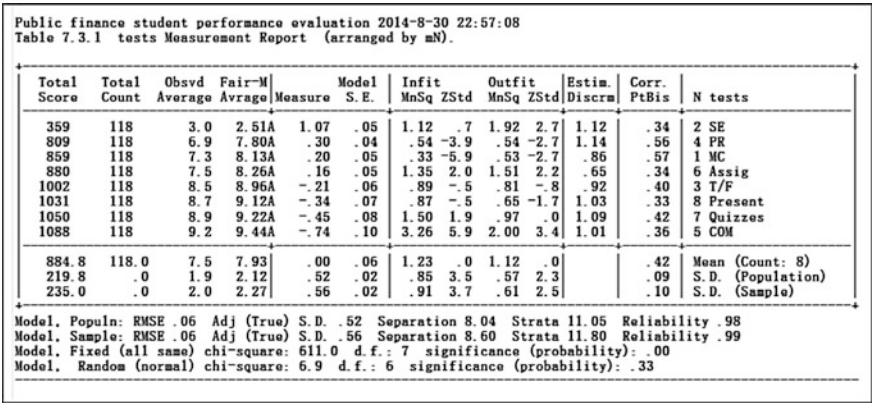
average. Foreign students (numbered 2, 6–8, 65–114, 118) had diverse performances. A few of them were retaking the course due to failure of the course before (numbered 2, 6, 7, and 8). These students had low performances, as expected. The rest of the foreign students had ranking of performances in the range from −0.29 to 2.30 logits. Hong Kong/Macau/Taiwan students (numbered 1, 3–5, 40–50, 113–114) concentrated in the range between −0.32 and 0.62 logits, which make this group of student the lowest achieving on average.

16.5.5 Test Facet Analysis

The eight tests of evaluation on the students showed different levels of difficulty to the students. The short essays had the highest difficulty and its distance from the less difficult items was also the greatest, which was 0.77 logits. The other tests had relatively low level of difficulty and the distance between each test was no more than 0.29. The short essay questions were designed to identify outstanding students in the study of the course; therefore its high level of difficulty was expected. The Presentation projects, online quizzes, and composition questions were designed to give students opportunity to earn marks simply by participation. Of course, such participation activities required the students to read, listen, speak out, analyze, write, and cooperate with others. The other tests such as multiple choice questions, problem-solving questions, and assignment questions were intended for students to perform comprehensive analysis using the knowledge learned in class. The levels of difficulty for these tests were higher than average, yet still achievable for most students.

Table 16.14 shows the results of analysis on tests. These tests had separation ratio of 8.04, reliability of separation ratio of 0.98, along with chi-square of 611.0, and probability of significance at 0.00, which indicated quite successful differentiation in terms of the tests.

Table 16.14 Tests measurement report



```
Public finance student performance evaluation 2014-8-30 22:57:08
Table 7.3.1  tests Measurement Report  (arranged by mN).

+---------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|          Model | Infit       Outfit     |Estim.| Corr.  |                    |
|  Score   Count  Average Avrage|Measure   S.E.  | MnSq ZStd   MnSq ZStd|Discrm| PtBis  | N tests            |
|-------------------------------+----------------+------------------------+------+--------+--------------------|
|   359     118       3.0   2.51A   1.07    .05  | 1.12    .7  1.92  2.7| 1.12 |    .34 | 2 SE               |
|   809     118       6.9   7.80A    .30    .04  |  .54 -3.9   .54 -2.7| 1.14 |    .56 | 4 PR               |
|   859     118       7.3   8.13A    .20    .05  |  .33 -5.9   .53 -2.7|  .86 |    .57 | 1 MC               |
|   880     118       7.5   8.26A    .16    .05  | 1.35  2.0  1.51  2.2|  .65 |    .34 | 6 Assig            |
|  1002     118       8.5   8.96A   -.21    .06  |  .89  -.5   .81  -.8|  .92 |    .40 | 3 T/F              |
|  1031     118       8.7   9.12A   -.34    .07  |  .87  -.5   .65 -1.7| 1.03 |    .33 | 8 Present          |
|  1050     118       8.9   9.22A   -.45    .08  | 1.50  1.9   .97   .0| 1.09 |    .42 | 7 Quizzes          |
|  1088     118       9.2   9.44A   -.74    .10  | 3.26  5.9  2.00  3.4| 1.01 |    .36 | 5 COM              |
|-------------------------------+----------------+------------------------+------+--------+--------------------|
|   884.8   118.0     7.5   7.93|    .00    .06  | 1.23    .0  1.12    .0|      |    .42 | Mean (Count: 8)    |
|   219.8      .0     1.9   2.12|    .52    .02  |  .85   3.5   .57   2.3|      |    .09 | S.D. (Population)  |
|   235.0      .0     2.0   2.27|    .56    .02  |  .91   3.7   .61   2.5|      |    .10 | S.D. (Sample)      |
+---------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .06  Adj (True) S.D. .52  Separation 8.04  Strata 11.05  Reliability .98
Model, Sample: RMSE .06  Adj (True) S.D. .56  Separation 8.60  Strata 11.80  Reliability .99
Model, Fixed (all same) chi-square: 611.0  d.f.: 7  significance (probability): .00
Model,  Random (normal) chi-square: 6.9  d.f.: 6  significance (probability): .33
-----------------------------------------------------------------------------------------------------
```

16.6 Conclusions

When students of the same course came from various regions all over the world, giving satisfactory lectures and evaluations is a challenging task. The analysis on the model fit statistics, student performances, test measurements, and item difficulty revealed that the system of evaluation had achieved satisfactory results. The system aimed to give students more opportunities in the study process, in the forms of both self-regulated study and cooperative activities. Judging from the results of the evaluation, the tests that were designed to be easy-achieving simply by participation and frequent practice and good cooperation gained good results. Students with ordinary abilities liked this kind of exercises since they could utilize their specialties in oral speech, team cooperation, and online exercises. The difficult parts of the tests were given in written exam forms such as short essays and multiple choice questions for students with stronger academic backgrounds and better abilities. The overall levels of difficulty match the composition of students with different backgrounds and abilities.

For further analysis, it is noticed that this evaluation system could not give students sufficient choices in terms of difficulty of questions within each test. Therefore, each type of test could be fine-tuned to give various levels of difficulty, so that students could be enjoying more different types of evaluation.

In policy suggestions to universities that seek to admit more international students, setting a practical standard of admission, offering courses, and conducting evaluations for diversified students would be vital in future success of education.

References

Caballé, S., Daradoumis, T., Xhafa, F., & Juan, A. (2010). Providing effective feedback, monitoring and evaluation to on-line collaborative learning discussions. *Computers in Human Behavior, 2010*. doi:10.1016/j.chb.2010.07.032.

Chen, C.-S., & Zhou, P. (2009). A correlation study on time management, self-regulated learning and general self—efficacy of collage students. *Journal of Xuzhou Normal University (Philosophy and Social Sciences Edition), 35*(3), 131–136.

Chen, Q., & Chen, J. (2010). A survey of non-English Majors' self-study. *Journal of Guizhou University for Nationalities (Philosophy and Social Science), 2010*(3), 187–191.

Demetriadis, S. N., Papadopoulos, P. M., Stamelos, I. G., & Fischer, F. (2008). The effect of scaffolding students' context-generating cognitive activity in technology-enhanced case-based learning. *Computers & Education, 51*(2008), 939–954.

China Association for International Education (2013). 2013 statistics on Foreign students into China from China Association for International Education. http://www.cafsa.org.cn/index.php?mid=6.

Gao, Q. (2012). *The Measurement of Foreign students' ability of writing in Chinese with a many-facets Rasch model*. Paper presented at the 2012 International Symposium on Information Technology in Medicine and Education.

Gokhale, A. A. (1995). Collaborative learning enhances critical thinking. *Journal of Technology Education, 7*(1), 22–30.

He, L., & Zhang, J. (2008). Investigating the reliability of CET-SET using multi-facet Rasch model. *Modern Foreign Languages (Quarterly), 31*(4), 388–398.

Hines, P., Oakes, P. B., Corljzy, D., & Lindell, C. O. (1998). Crossing boundaries: Virtual collaboration across disciplines. *The Internet and Higher Education, 1*(2), 131–138.

Horsburgh, M., Lamdin, R., & Williamson, E. (2001). Multi professional learning: The attitudes of medical, nursing and pharmacy students to shared learning. *Medical Education, 1*(35), 876–883.

Jeffrey, L. M. (2009). Learning orientations: Diversity in higher education. *Learning and Individual Differences, 19*(2009), 195–208.

Li, Y., & Xu, L. (2010). The application of management strategy in self-study by non-English majors. *Journal of Chengdu University of TCM (Educational Science Edition), 12*(2), 43–44.

Lytras, M. D., & Ordóñez de Pablos, P. (2009). *Social web evolution. Integrating semantic applications and web 2.0 technologies*. Hershey: IGI-Global.

Moller, L., Huett, J., Holder, D., & Young, J. (2005). Examining the impact of learning communities on motivation. *The Quarterly Review of distance Education, 6*(2), 137–143.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge. Acting with technology series*. Cambridge, MA: MIT Press.

Sun, H. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English Test for non-English major graduates. *Chinese Journal of Applied Linguistics Bimonthly, 33*(2), 87–102.

Wang, L., & Huang, X. (2013). Applying the many-facet Rasch model to evaluate college students' school subject competence. *Examinations Research* (General No. 35), 41–50.

Wang, Q., Woo, H. L., & Zhao, J. (2008). Investigating critical thinking and knowledge construction in an interactive learning environment. *Interactive Learning Environments, 17*(1), 1–10.

Chapter 17
Correlation Analysis on Influencing Factors of English Writing and Non-English Major Tibetan College Students' English Writing Development

Baina Xia

Abstract This study aims to research on the correlations between the five aspects in English writing, viz., content, organization, vocabulary, language use, and mechanics, and the three important factors that affect students' writing quality, viz., learners' English proficiency, writing strategy, and feedback. The researcher compared statistically 108 non-English major Tibetan college students' writing score in their first year of college. During this academic year, they were required to take the English writing course; the evaluating scheme of the tests and compositions stayed the same. A 118 English compositions of eight students in this year were investigated and analyzed to reveal the development of the five aspects in their writing. Five of these eight students were interviewed in the period of their writing course so that more influencing factors can be discovered. By the end of their study in the first year, questionnaires were distributed to the 108 students. The data from their questionnaire and five aspects of their writing score were analyzed statistically. The analysis of the above-mentioned data show that: (1) English writing quality of these learners improved prominently after study in college for 1 year; (2) four of the five aspects in writing—content, organization, vocabulary, and language use—have shown improvement of different levels with exceptional but explainable changes, and the interviews can explain why the improvement is not persistent, while mechanics does not have a stable pattern of development; (3) in most cases, content, organization, vocabulary, and language use can be predicted by English proficiency and writing strategy, but cannot be predicted by feedback; mechanics cannot be predicted by these three influencing factors. In the end, implications of this study and suggestions for further studies are proposed. If positive factors can be reinforced and negative ones deleted, both English writing's learning and teaching can be more effective.

Keywords English writing · English proficiency · Writing strategy · Feedback · Content · Organization · Vocabulary · Language use · Mechanics

B. Xia (✉)
Faculty of Foreign Studies, Tibet Nationality College, 712082 Xi'an,
Shaanxi, China
e-mail: xiabaina@hotmail.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_17

17.1 Introduction

English writing is one of the four main domains in English learning. For decades, people who learn and teach English labor to explore better ways of improving English writing especially EFL English writing. Former researches and studies have covered many aspects of EFL writing, yet the gap between EFL writing's development in a certain period and its influencing factors are yet to be bridged. This research attempts to investigate non-English major Tibetan students' writing performance change after experiencing one year's writing course in college. The second attempt of this research is to analyze how these students' composition score change statistically and how their writing develop in the aspect of content, organization, vocabulary, language use, and mechanics. Furthermore, the researcher discovers to what degree the three important factors, viz., English proficiency, feedback, and writing strategy, influence the development. In the light of the results, students may improve their English writing, and teachers improve their teaching in a reasonable situation. After a year's study, if the scores made difference, the changing pattern of the score could be presented. A further step was to discover how the five evaluating aspects of EFL writing developed during this year so that detailed patterns of development were detected and perceived. From the result of statistical analysis, this study reveals the connection between the three main influencing factors and the development of EFL writing's five aspects.

17.2 Procedure

After entering university, 108 Tibetan students were enrolled as non-English majors, but were assigned in the same class level. Although their English writing was tested in the entrance exam, they did not have English writing course in middle school. An English writing test was arranged immediately after they entered university, by the end of the first and second term. Eight students' 118 pieces of writing assignments were selected. During the time between these two semesters, they were asked to finish a questionnaire concerning the influencing factors of their writing; and several volunteers were interviewed to work on the refinement of the questionnaire as well as covering as many influencing factors as possible.

The questionnaire has totally 50 items. The first 20 items check students' English proficiency with four for the five evaluating aspects each. Item 21 to item 40 are planned to check their English writing strategies also with four items for each five aspects. Item 41 to item 50 are designed to check the feedback with two for the five each. Students are expected to fill the questionnaire by ticking a number behind each item which scaled from one, which means false or never, to five, which means true or always. During the process of designing the questionnaire, five students were voluntarily interviewed on what had been influencing their English writing proficiency.

17.3 Results Analysis

17.3.1 108 Students' Writing Scores

The 108 Tibetan college students took three writing tests at their entrance of college, by the end of the first semester, and at the end of their second semester, respectively. The statistics showed that the results of these three tests are significantly different in statistics.

The comparatively lower writing score in the first test has indicated that students' writing score rose after their study in college writing course. In the second and third test, test 2 has the higher mean score, more students have reached level of good or excellent, however, in test 3, and more students have scores higher than mean score. Most of the 108 students' writing reached the top level after one semester's study in writing course; however, by the end of the second semester, their writing scores as a group are more stable. The students have shown improvement in their writing quality after one year's study in college.

17.3.2 Individual Students' Writing Development

While the author followed the eight individual Tibetan students' compositions, scores of content, organization, vocabulary, language use, and mechanics in their writing were separately recorded and considered.

First, as far as content score is concerned, they all show unsteady movements, yet seven of the eight students have shown different level of increase, and by the end of the second semester, the content score have all reached the level of very good to excellent; only student 1 does not have significant increase in content score, but his content score move up and down between the level of good and excellent. Student 1 and 3 have very low content score in composition 5; Student 5 in composition 6; Student 2, 4, and 6 in composition 7, and student 4 also show very low content score in composition 8; yet Student 7 shows very high content score in composition 6.

Second, the first five students' organization scores do not show significant improvement prominently. But organization scores in their writing all move slowly from average at the very beginning to very good by the end of the second semester. While Student 6, 7, and 8 have shown greater development on the organization score which have moved from level of fair in first composition to very good by the 15th composition. Student 1 and 3 have very low organization score in composition 5; Student 5 in composition 6; Student 2, 4, and 6 in composition 7, and student 4 also show very low organization score in composition 8.

Third, except Student 4 and 8, the other six students' showed little prominent change in their vocabulary score of writing. The score just bounced from good to very good, or average to good. Student 4's vocabulary score is not stable during the 15 compositions, even had very low scores in the middle four compositions, but at last, the score moved steadily to the level of very good. Student 1's vocabulary score, like his content score and organization score, is very low compared with his other compositions; and Student 2's vocabulary score is very low, too, as is shown in her content and organization score. While Student 8's vocabulary score move from average to very good at the first eight compositions, but moved down to average level in the compositions followed.

Fourth, in Student 2, 3, 4, 7, and 8's writing, the language use score have all shown very unsteady improvement from level of average to good, then to very good. While Student 1, 5, and 6 do not have significant improvement in language use score through the 15 compositions, student 1 and 3 show particular low language use score like content, organization, and vocabulary score in composition 5; and Student 2 have very low language use score in composition 7.

Fifth, Only Students 2, 4, and 7 have low mechanics score in some compositions, the other five students have low score in mechanics before the 6th composition, but keep stably high mechanics score after that.

The results of Pearson correlation test and summary of interviews have proved that the particular high or low scores are not exceptional, since from the interviews, it is found that students have preferences in types and topics of compositions. The four aspects, content, organization, vocabulary, and language use, have significantly high correlations with one another. The unpredictable mechanics score in the individual student's compositions can also be explained from the results, which shows that mechanics score have no significant correlations with content, organization, and vocabulary score, and very small correlations with language use score.

17.3.3 Influence of English Proficiency, Writing Strategy, and Feedback on Students' Writing

The data indicates that English major students' English proficiency, writing strategy, and writing feedback statistically have significant correlations with their writing and five different aspects of writing: content, organization, vocabulary, language use, and mechanics. The three factors can be used as prominent predictors to predict students' writing total score, and its five aspects' score, and in English writing score, 80–97 % of the standard deviation can be predicted by the questionnaire's result which shows students' different English proficiency, writing strategy, and their writing feedback. How can the three factors affect students' English writing and the five aspects consisted?

First, former research has proved that English proficiency can influence strategy use (De Larios et al. 1999; Sasaki 2000) and self-confidence of learners (Cheng 2002). But the present study discovered more detailed influence of English proficiency on EFL writing quality. Students' English proficiency on their writing content, organization, and vocabulary have very prominent and significant correlations with their writing scores of content, organization, and vocabulary, in which the highest correlation is between English proficiency on vocabulary and vocabulary score in writing; and the lowest correlation is between English proficiency on content and content score. English proficiency on language use also has significant correlation with language use score, but it has no contribution to predict language use score. Students' English proficiency on mechanics does not have significant correlation with mechanics score.

Second, writing strategy on English writing's content, organization, vocabulary, and language use statistically also have prominent correlations with students' writing scores on content, organization, vocabulary, and language use respectively, which has further supported former research (Long 1983; Santangelo et al. 2007; Sexton et al. 1998; Xiao 2008). However, the role of writing strategy in English writing is also further explained in this research since it has been proved in the present study that writing strategy on content and content score bear the highest correlation; and writing strategy on language use and language use score have comparatively lower correlation; writing strategy on vocabulary is a negative predictor of vocabulary score. Still, writing strategy on mechanics has no statistically significant correlation with mechanics score.

Third, many research have proved positive influence of feedback on English writing (Chandler 2003; Enginarlar 1993; Ferris 1995, 1997; Miao et al. 2006; Wang and Dong 2010). However in the present study, results have shown that English majors' writing feedback on organization and language use have prominently significant correlations with organization score and language use score in writing, however, writing feedback on organization is the negative predictor of organization score. While, writing feedback on content, vocabulary, and mechanics do not have statistical correlation with content score, vocabulary score, and mechanics score in writing.

To sum up, the three influencing factors of English writing, English proficiency on content, organization, and vocabulary have strong and significant correlations with content, organization, and vocabulary in writing, and they are prominent predictors of these three aspects in writing; English proficiency on language use cannot be used to predict language use score. Writing strategy on content, organization, and language use are significant predictors of content, organization, and language use; writing strategy on vocabulary can also be used to predict vocabulary score, but the more strategy on vocabulary students grasp, the lower vocabulary score they may get. Unlike results from previous researches which indicate positive influence of feedback on EFL writing, in this research, more writing feedback on language use can predict higher language use score, and more feedback on organization may lower organization score. Other than that, feedback does not have correlations with content or vocabulary. Mechanics do not have correlations with English proficiency, writing strategy, or feedback, and cannot be predicted by the three factors.

Besides English proficiency, writing strategy, and feedback, there are other factors that can affect EFL learner's English writing. Learner's instructional background, preference in styles and topics, and learner's attitude toward writing course are the three main factors revealed in the interviews.

17.4 Implications

Research findings indicate that the 108 non-English major Tibetan students' writing course had positive influence on their writing quality. Hence, the following pedagogic implications are proposed:

First, it is necessary to put more impact on teaching students how to properly use the words they learned in writing instructions. Although many students put their focus on accumulating new words and phrases in their study, their vocabulary in writing did not improve in two semesters' study, besides, writing strategies have negative influence on vocabulary in writing. This means the new words and phrases they accumulated were not applied in their writing. Since students do not take the initiative in writing in English after class, output practice of vocabulary learning should be enhanced in class.

Second, EFL English learners should notice and adopt writing strategies in English writing to improve their writing quality. English proficiency and writing strategy are important influencing factors of students' English writing. If students want to have better English writing, they should focus not just on writing, but also on other aspects of English, reading, speaking, and listening. Students who read more English or Chinese literature can be more creative and ideas are better illustrated in there writing. In speaking class, writing and speaking can be mutually beneficial, while listening is also an effective input in the process of English acquisition. Learning writing strategy should be systematic and draws more attention in writing instructions.

Third, Peer feedback and self revision are factors even more difficult to control, not only teachers, but learners themselves should build a filter for the feedback since not all feedback is useful or even correct. Feedback does not influence students' English writing as much as the English proficiency and writing strategy, it even has negative effect on organization in writing. The initial intention of feedback is to make learners' English writing better, however, the amount and types of feedback in class should be well adjusted by teachers. Sometimes, errors or mistakes should only be underlined and pointed out; while in other situations, they should be corrected or even with subtle illustrations.

17.5 Suggestions for Further Studies

Based on the findings, English major freshmen's improvement on writing and the impact of the three influencing factors are evident. Advice for teaching and learning English writing are elicited. Moreover, the findings have illustrated many related questions which may provide directions on future studies of second language learning and teaching.

First, research can be conducted between different study groups and in multi-levels. Students of different classes or grades have different learning environment which may cause distinctive writing output. If time and resource permit, research can also be conducted among learning groups from different colleges. Study has shown an increase in L2 writing anxiety as time of study increases (Cheng 2002), there might be other factors changing with learning time and conditions. Only college freshmen participated in this research, learners of other levels, for instance, senior college students or people who have graduated from college, may have different developing patterns or influencing factors in their English writing.

Second, studies can be carried out more intensively with more influencing factors added and considered. There are more than three factors that can affect students' English writing (Bachman 1999; Cheng 2002; Spolsky 1989; Williams 2007), the more factors revealed, the more guidance the studies may offer to the future EFL teaching and learning.

Third, with the increasing popularity of web-based education, researches can be conducted in internet environment (Wible et al. 2001), since computer-mediated instruction has already been put into practice in foreign language education. If web is adopted in EFL writing course, teachers may be free from the laborious workload of examining students' writing. More compositions can be collected and evaluating standard can be strictly followed since factors affecting the inaccurate assessment of students' compositions are reduced. Hence clearer patterns of their writing's development might be discovered.

References

Bachman, L. F. (1999). *Fundamental considerations in language testing*. Shanghai: Shanghai Foreign Language Education Press.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), 267–296.

Cheng, Y. (2002). Factors associated with foreign language writing anxiety. *Foreign Language Annals, 35*(5), 647–656.

De Larios, J. R., Murphy, L., & Manchon, R. (1999). The use of restructuring strategies in EFL writing: a study of Spanish learners of English as a foreign language. *Journal of Second Language Writing, 8*(1), 13–44.

Enginarlar, H. (1993). Student response to teacher feedback in EFL writing. *System, 21*(2), 193–204.

Ferris, D. R. (1995). Student reactions to teacher response in multiple-draft composition classrooms. *TESOL Quarterly, 29*(1), 33–53.

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly, 31*(2), 315–339.

Long, M. H. (1983). Does second language instruction make a difference? A review of research. *TESOL Quarterly, 17*(3), 359–382.

Miao, Y., Badger, R., & Zheng, Y. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.

Santangelo, T., Harris, K. R., & Graham, S. (2007). Self-regulated strategy development: A validated model to support students who struggle with writing. *Learning Disabilities, 5*(1), 1–20.

Sasaki, M. (2000). Toward an empirical modal of EFL writing processes: An exploratory study. *Journal of Second Language Writing, 9*(3), 259–291.

Sexton, M., Harris, K. R. & Graham, S. (1998). Self-regulated strategy development and the writing process: Effects on essay writing and attributions. *Exceptional Children*, *64*(3).

Spolsky, B. (1989). Communicative competence, language proficiency, and beyond. *Applied Linguistics, 10*(2), 138–156. doi:10.1093/applin/10.2.138.

Wang, W., & Dong, Y. (2010). The justification of teacher-guided error correction of Chinese college students' English writing. *Chinese Journal of Applied Linguistics (Bimonthly), 33*(3), 63–75.

Wible, D., Kuo, C., Chien, F., Liu, A., & Tsao, N. (2001). A web-based EFL writing environment —integrating information for learners, teachers, and researchers. *Computer & Education, 37* (3), 297–315.

Williams, J. (2007). *Teaching writing in second and foreign language classrooms*. Beijing: World Publishing Corporation.

Xiao, L. (2008). Exploring a sociocultural approach to writing strategy research: Mediated actions in writing actions. *Journal of Second Language Writing, 17*(4), 217–236.

Chapter 18
An Empirical Case Study of TEM-8 Test Reliability

Jianfang Xiao

Abstract Tests are a common assessment method and an effective way to examine language learning. The Test for English Majors Band 8 (TEM-8), a large-scale and high-risk test administered by Ministry of Education China, has been carried out for twenty-three years. For testing quality, reliability, and validity are two extremely important criteria and a test must have reliability prior to validity. The author conducted a research on 10 language study periodicals included in CSSCI and found that empirical studies of language testing reliability had rarely been reported. This study analyzed the reliability of the TEM-8 test papers from 2009 to 2013 based on analyses of the scores of 20 students randomly selected from a class who did the 2009–2013 TEM-8 test papers in mock tests. The results show that (1) The 5 years' TEM-8 test papers are of high reliability; (2) As for the subjective items, especially the translation and writing part, there is a large difference between the students' achievement at various tests, and a big difference in difficulty values. Hence, reliability of this part is lower. (3) In terms of objective items, the students maintained relatively high stability and consistency in achievement and the reliability of these items is higher.

Keywords TEM-8 test papers · Reliability · Validity · Levels of difficulty

The writing of this paper is financially supported by Research Center for English Testing and Assessment, Guangdong University of Foreign Studies, 'the present status of and future improvement on college English teachers' teaching abilities in Guangdong Province' approved by Guangdong Provincial Office for Social Sciences Planning (GD12XJY02) and the project 'A comparative study of foreign language education policies in European Union countries' approved by China's National Office for Education Sciences Planning (DDA130230).

J. Xiao (✉)
School of English and Education, Guangdong University of Foreign Studies,
Guangzhou, China
e-mail: xiao3296888@aliyun.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_18

18.1 Introduction

Tests are a common assessment method and an effective way to examine language learning. The Test for English Majors Band 8 (TEM-8), a large-scale and high-risk test administered by Ministry of Education, China, has been carried out for twenty-three years. Its test papers are prepared by the English Panel of China's National Supervisory Committee of FLT in Higher Education based on the requirements of "English Teaching Syllabus for English Majors" and the test is conducted nationwide by the Steering Committee Office.

Tests, especially large-scale high-risk ones, must have a number of criteria including reliability, validity, and feasibility (including the levels of difficulty, discrimination, operability, repeatability, beneficial washback, scores interpretability, and economic affordability), of which reliability and validity are extremely important. If a test loses reliability and validity, the other criteria will be out of the question (Gui 1986). Reliability is also called dependability or consistency. Bachman and Palmer (1996) maintain that reliability of a test should be sufficient to answer this question: how many of the errors committed by a test-taker are caused by factors beyond his/her ability? Or put it in another way, to what extent is an examinee's score reliable or trustworthy? The language testing expert Brown (2006) also defines reliability as the degree of consistency or stability of test results. China's applied linguistics community did not attach much importance to the application of and research into the test reliability and validity theory until the 1990s. In recent years, an increasing number of experts and scholars have conducted in-depth research in the field from different aspects. This paper reviews the major researches on the reliability and validity of English testing in China in the past 11 years and finds such researches in China focuses on validity, particularly content validity, and empirical research on test reliability from the perspective of test takers is scanty. Thus, this study intends to carry out a reliability study of the 2009–2013 TEM-8 papers.

18.2 Current Status of Research on Language Test Reliability and Validity in China

The present author used "English test validity" and "English test reliability" as search key words to retrieve research documents published on China's CNKI from January 2004 to December 2014. A total of 343 papers were published during the period.

The data shows the concern for research on English test validity and reliability is overall on the upward trend (see Table 18.1). The author searched 10 foreign language research and teaching periodicals (*Foreign Language Teaching and Research, Modern Foreign Languages, Foreign Languages, Foreign Languages World, Foreign Languages and Foreign Language Teaching, Foreign Languages in China, Foreign Language Teaching, Journal of PLA Foreign Languages*

Table 18.1 Number of research papers on reliability and validity on CNKI 2004–2014

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Number	23	16	23	28	26	31	26	30	35	45	60

Table 18.2 Number of papers on reliability and validity in 10 Foreign language research and teaching journals included in CSSCI 2004–2014

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Number	1	4	6	2	2	6	4	4	0	1	0

Institute, Foreign Languages Journal, and Foreign Language Research) and browsed the papers on test validity and reliability, 30 articles in all. On the whole, the research quantity bobbed up and down (see Table 18.2). It peaked in 2006 and then quickly fell back, and then in 2009 it reached another peak and began to decline after 2012.

Based on the specific content of research, the 30 articles fall into 3 categories, of which 21 are on validity of language testing, accounting for 70 % of the total; 5 on reliability study, accounting for 16.7 %; and the remaining 4 covers both, accounting for 13.3 % (see Table 18.3).

From Table 18.3, it can be seen the validity of oral English tests got the most attention of researchers among the various items. Speaking as an output test item can directly and effectively test the examinee’s oral expression ability. If the design is reasonable, it can have positive backwash effect on language teaching. But in large-scale tests, it is difficult to ensure the consistency and accuracy of the examinees’ scores. For this reason, oral test has not been included in large-scale language tests so far. But the development of globalization has been raising the requirement for oral English ability, and for the effective measurement of this ability. A number of experts including Bao (2009), Guo et al. (2012), Wang Haizhen and Wen Qiufang have conducted research on the reliability and validity of oral testd (Wen 2009). While the study of the validity and reliability of subjective test questions in language tests has got increasing attention, the validity of objective test questions has also come under the scrutiny of related scholars. He Yongbin did research on listening test (2005), Xu and Zhang on grammar and vocabulary test (2004), Guo on cloze test (2010), and Zou Shen and Yang on reading comprehension test (2011). With the rapid development of computer and network technology, language test media are changing gradually. It is the wave of the future that computer-based language testing (CBLT) will replace paper-and-pen-based language testing (PBLT) and research on the reliability and validity of computer-based language testing is getting increasing attention of experts (Li and Wen 2009).

Analysis of the status quo of research in this field in China reveals some inadequacies. First the research is not balanced. Oral test, influenced by subjective scoring, has relatively low reliability which in turn affects its validity. It thus has got the greatest concern of language testing experts, with the papers published in this aspect being the greatest in number. The reliability and validity of other test items,

Table 18.3 Content-based categorization of papers on reliability and validity carried in the above-mentioned 10 journals 2004–2014

Categories	Main content	Number of papers		Proportion (%)
On validity	On the validity of spoken English testing	4	21	70.0
	On validity theory and research review	3		
	Computer-based test validity research	3		
	On cloze test validity	2		
	On validity of grammar and vocabulary testing.	1		
	On validity of interpretation testing	1		
	On validity of reading testing	2		
	On validity of writing testing	1		
	On validity of general knowledge about major English-speaking countries testing	1		
	On validity of post-listening dialogue completion	1		
	On validity of teacher-designed tests	1		
	On validity of listening tests	1		
On reliability	On rater reliability	4	5	16.7
	On reliability calculation models	1		
On both reliability and validity	On reliability and validity of oral tests	2	4	13.3
	On overall reliability and validity	1		
	On reliability and validity of computer-based tests	1		

however, cannot be ignored, for only when the reliability and validity of each part of a test is raised can the overall reliability and validity of a test be improved. Second, currently insufficient attention has been paid to the reliability and validity of formative assessment and teacher-designed end-of-term tests. Formative assessment and teacher-designed final exams have direct backwash effects on teaching, and therefore deserve more attention of researchers. Third, it can be seen from the published papers that scant attention has been paid to reliability of tests. There are only five papers in this aspect and one on the calculation models of reliability and the remaining four are on reliability from the angle of raters. Studies have found that the TEM-8 is high in validity; it sets real tasks and can test the comprehensive language ability of the examinees and thus can play a positive role in promoting teaching. For such large-scale, high-risk tests as TEM-8, it is necessary to conduct reliability research to ensure their fairness.

18.3 Research Design

18.3.1 Research on Reliability of Language Testing

Research on test reliability generally can be divided into two categories: scorer or rater reliability and test reliability. At present, published research on reliability of language testing in China is basically carried out from the rater angle. Research has found out that the examiner's severity, the difficulty level of the task, scoring criteria, and the scale factors are all likely to produce a certain amount of measurement error, which leads to the difference of scores of the candidates (He and Zhang 2008; Sen Zhang and Yu 2010; Liu 2010; Li 2011; Xiao 2011). This paper intends to study the TEM-8 score differences from the angle of test-takers.

18.3.2 Research Questions

This paper attempts to study the 2009–2013 TEM-8 papers and explore the reliability of the TEM-8 test papers from the angle of examinees with a view to answering the following three questions:

(1) Is there consistency and stability in the papers through the years?
(2) Is there consistency between the objective questions and subjective questions?
(3) Is there consistency and stability between the students' 2014 test results and their scores for the 2009–2013 test papers?

18.3.3 Research Materials and Subjects

18.3.3.1 Research Materials

The research materials for this study are the TEM-8 test papers of 5 consecutive years from 2009 through 2013. The proof-reading and error correction, writing and translation (English to Chinese and Chinese to English translation) are subjective questions and the rest take the multiple-choice form.

18.3.3.2 Subjects of the Study

The subjects in this study are 20 students (seniors) in a regular class of English majors which is a part of the 108 English majors of Grade 2010 of English Department in University G. The regular classes were formed upon enrollment taking into consideration such factors as students' sex, regional background, total scores, and scores on English in the College Entrance Exam. The selection of a regular class as the subjects of study could preclude the factor of imbalance. The selection of the seniors as study subjects lay in the fact that they were to take the TEM-8 test soon and were psychologically prepared to take a number of mock TEM-8 tests. They willingly pitched into test their English proficiency and accumulate testing experience. Thus the sampling is of great significance.

18.3.4 Research Procedures

The 2014 TEM-8 was to take place on 22 March, so the mock tests were scheduled in the second half of the first semester and the first one-third of March in the second semester of their fourth year. There were altogether five mock tests, taking place every Saturday and each lasting 185 min. The tests took place just as it is done in a regular TEM-8 test. The test venue was a language laboratory where the audio equipment was adjusted in advance and students were given answer sheets. The students were told the scores of the five mock tests would be taken into their final score for the "Advanced English" course and they were to do the test in a serious manner. In order to ensure the test reliability, the students were not told they were doing used real TEM-8 test papers which were not used in their chronological order but in the order of 2010, 2009, 2012, 2011, and 2013. Meanwhile, in order for the students to benefit from the testing, each test paper done by the students was explained to the students on the following Thursday afternoon (the students were free during this time), which was warmly welcomed by the students.

18.3.5 Collection of Research Data

After the papers were collected, the objective questions were rated by a scoring machine in a unified way. In order to reduce the effect of scoring bias on the reliability of the test scores, the subjective parts were scored by teachers of the Advanced English course, who before beginning their work, perused the scoring criteria. See Table 18.4 (scores for translation and writing) and Table 18.5 for the statistics. In Table 18.5, the results for the 2014 test were official results based on the students' actual TEM-8 test.

18.4 Results and Analysis

18.4.1 Consistency and Stability in the 6 Years' (2009–2014) TEM-8 Test Papers

In order to observe the consistency and stability of the 6 years' TEM-8 test papers, the SAS (Statistic Analysis System) statistical analysis software was used to analyze the total scores of the 20 students in their 2009–2014 papers. See Fig. 18.1 and Table 18.6 for results.

As can be seen from Fig. 18.1, the fluctuations of the six curves are consistent, indicating the students' overall consistency and stability in the six tests. Table 18.6 shows that students No. 1, 2, 4, 5, 7, and 13 were almost in the forefront at all times and the ranking of students No. 1, 2, 4, 6,7, 8, 18, and 19 in the various tests did not

Table 18.4 20 students' scores on subjective questions (translation and writing) of TEM-8 test papers 2009–2013 year

Year	2009		2010		2011		2012		2013	
Type of question	Translation	Writing	Translation	Writing	Translation	Writing	Translation	Writing	Translation	Writing
Student										
S1	8	14.5	16	15	17	15	16	14.5	12	15
S2	8	15	17	16	19	14.5	17	15	11	16
S3	7	11	17	14	12	14	9	13	9	14.5
S4	12	17	17	16	17	15	19.5	16	19	15
S5	5	14.5	18	14.5	15	14.5	16	14	13	14
S6	9	12	15	15	12	14	15.5	14.5	12	14
S7	8	15	17	15	17	15	18	16	13	145
S8	11	13	15	14.5	15	14.5	11	14	11	15
S9	7	13	16	14.5	13	14	12	14.5	7.5	14.5
S10	9	14	13	15	13	14.5	12	15	11	15
S11	13.5	14	15	14	13	14.5	16	13	8	15
S12	9	13	12	11	15	14	15	14	15	14
S13	9	14.5	17	12	17	15	15.5	14.5	8	16
S14	7	13	13	11.5	9	13	15	14	5	14.5
S15	12	11	17	14	16	14.5	16	14.5	11	14
S16	4	14	17	14.5	15	14	15.5	14.5	8	14.5
S17	5	14	15	14.5	7	14	13	14	4	14.5
S18	3	13	11	14.5	17	14	11	14	3	14
S19	8	12	16	14	18	13	13	13	19	13
S20	11	14.5	15	15	13	14.5	13	16	12	15

Table 18.5 20 students' scores on the objective questions (obj) in 2009–2013 TEM-8 test papers and their total scores and their scores on the 2014 TEM-8 test

Year	2009		2010		2011		2012		2013		2014
Types	Obj	Total	Obj	Total	Obj	Total	Obj	Total	Obj	Total	Total
Students											
S1	31	61	35	72.5	37	76	36.5	72.5	37	72	75
S2	35	66.5	29	71.5	33	71.5	33	70.5	36	70.5	76
S3	31	56	19	52	33	65.5	33	62	35	66	69
S4	36.5	73.5	28.5	65.5	35.5	74.5	35.5	77	40	81.5	79
S5	34	62	33	70	37.5	73.5	36	71	42	75	70
S6	29.5	57.5	29.5	65.5	35.5	67.5	26.5	61.5	35.5	66.5	70
S7	33	64	32	69	35	71	29.5	69	36	69	73
S8	29	58	30.5	65.5	34.5	66.5	30.5	62	33.5	66.5	72
S9	32	58	28	63.5	29.5	62.5	35.5	68.5	36.5	66	71
S10	29.5	59.5	26	60	33.5	67.5	29.5	63.5	35.5	67.5	72
S11	24	55.5	26	59	34.5	68.5	32.5	66.5	31.5	62.5	71
S12	31	60.5	31	60.5	30.5	64.5	28	61	33	67	69
S13	31	62.5	29	66	35.5	74	30.5	66	35	67	74
S14	26	52	29.5	57	29.5	58.5	24.5	57	32	60	58
S15	29	57.5	28	63	31	66.5	28.5	63.5	36	67	71
S16	30.5	54.5	29	64.5	35	69	26	61.5	36.5	66.5	66
S17	35	63	29	64.5	35.5	63.5	29.5	65.5	32.5	59.5	71
S18	29	53	23	53	27.5	63	26	56	27.5	51.5	63
S19	27	53	29	64.5	28.5	63.5	29.5	61.5	32	60	64
S20	26	57.5	28.5	68.5	35	70	31.5	67.5	37	72.5	74

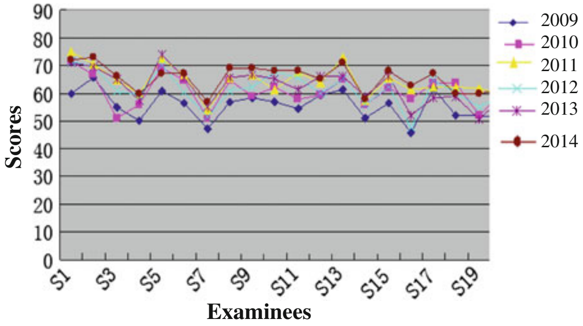


Fig. 18.1 20 students' scores on 2009–2014 TEM-8 test papers

change much. The 20 students' scores for their various mock tests were consistent with their score for the 2014 formal TEM-8 test and all the test results were highly correlated. This shows there is consistency and stability between the TEM-8 takers' scores for the 5 years from 2009 to 2013 and the 2014 TEM-8 takers' results. The above analysis demonstrates that the TEM-8 test papers from 2009 to 2014 were of relatively high reliability.

Table 18.6 20 students' (S) respective places (P) in the 2009–2014 TEM-8 tests

Year	09	10	11	12	13	14
P/S	P	P	P	P	P	P
S1	7	1	1	2	4	2
S2	2	5	5	4	5	3
S3	15	20	14	13	14	15
S4	1	9	2	1	1	1
S5	6	2	4	3	2	12
S6	12	7	10	15	11	13
S7	3	3	6	5	6	5
S8	10	8	12	14	12	7
S9	11	15	18	6	15	9
S10	9	17	11	11	7	8
S11	16	18	9	8	16	10
S12	8	16	15	18	8	15
S13	5	6	3	9	9	4
S14	20	19	20	19	17	20
S15	13	14	13	12	10	11
S16	17	12	8	17	13	17
S17	4	10	16	10	19	14
S18	19	22	18	20	20	19
S19	18	11	17	16	18	18
S20	14	4	7	7	3	6

18.4.2 *Consistency Between the Subjective Items and Objective Items in the 2009–2014 TEM-8 Test Papers*

Although the six curves in Fig. 18.1 fluctuate consistently, there is a relative large difference between the highest and lowest curves. Language learning theory and practice tells us that a student's language performance cannot fluctuate so much in such a short period of time. To find out the cause, the following statistics were done carefully:

Table 18.7 shows that the students' TEM-8 scores jumped in the space of a few months from 59.25 for the 2009 test paper to 74.05 for 2014 test paper, with a difference of 14.80 points. The statistics in the table reveals the following causes: ① The translation part of the 2009 and 2013 papers were of a relatively high level of difficulty, with the difficulty value of the 2009 paper being 0.44 and the 2009 paper being 0.53. This is the main cause for the low scores of 2009 and 2013. ② Viewed from the discrete quantity, the range of the 2009 test paper was 21.5 and that of 2013 was 30. Too large a range is another reason for the low scores of 2009 and 2013. ③ Viewed from the statistical standard deviation, the standard deviation of the 2009 test paper was 6.179 and that of 2013 was 6.968; the degree of discreteness is rather too large. In brief, the 2009 and 2013 papers were overall too

Table 18.7 Statistics of 20 students' scores, test difficulty value, concentration magnitude, and discrete quantity in the 2009–2014 TEM-8 test papers

Items	Translation		Writing		Objective	Total score		
Year	Average	Difficulty	Average	Difficulty	Average	Average	Range	Standard deviation
2009	9.775	0.44	16.575	0.67	31.2	59.25	21.5	6.179
2010	17.225	0.78	14.425	0.72	29.1	64.05	20.5	5.591
2011	14.5	0.70	14.2	0.71	33.48	67.85	27.5	5.296
2012	14.55	0.73	14.45	0.71	30.55	65.18	21	6.418
2013	10.45	0.53	12.625	0.80	35	66.68	30	6.968
2014	14.6	0.63	14.6	0.68	33.35	74.05	21	4.942

difficult. To avoid the defect of too small samples, the average pass rates TEM-8 (total score 60 points or above) of the country from 2009 to 2014 were officially retrieved, which were 40.33 % (2009), 43.11 % (2010), 42.44 % (2011), 41.49 % (2012), 40.08 % (2013), and 42.76 % (2014), respectively. The results were consistent with this study's analysis of the difficulty of the papers. Table. 18.4 and 18.5 show that there was not adequate difficulty stability and consistency in the subjective parts of the TEM-8 test papers in the recent years.

18.5 Conclusions and Suggestions

Analysis of and research on 2009–2013 TEM-8 test papers and related data from the angle of examinees offer the following conclusions:

(1) Overall, there was consistency between the 2009 and 2014 TEM-8 test papers and therefore they were of a relatively high reliability. Figure 18.1 and Table 18.6 show the scores of the 20 examinees were in a highly consistent trend. What is particularly noteworthy is students No. 1, 2, 4, 5, 7, and 13 were almost in the forefront at all times, and the scores of students No. 1, 2, 4, 6, 7, 8, 18, and 19 did not fluctuate much. Moreover, the objective questions in the 2009–2014 papers were scientifically designed and the students' average scores on the objective test items ranged from 31.2 (2009) to 33.35 (2014), with a difference of only 2.15 points. The students' scores for the objective test parts show a high degree of stability and consistency.

(2) To some extent, the subjective questions in the TEM-8 test papers were not stable in difficulty levels and thus lack stability and consistency. From Table 18.6 it can be seen that the difficulty level of the translation parts in the 2009 and 2013 papers, with their difficulty value at 0.44 and 0.53, respectively, was obviously higher than that of other years' papers, which was likely the main reason causing the average scores for the translation part in the 2 years to be excessively low. Also, compared with other years, the writing part in the 2009 paper was more difficult, with the difficulty value being 0.67.

(3) The students' scores for the 2009–2013 TEM-8 papers used in the mock tests and those on the 2014 TEM-8 formal test showed consistency and stability, so the students' scores on the TEM-8 test were highly reliable. Table 18.6 shows the 20 students' scores on the 2009–2013 TEM-8 papers were highly consistent with their scores on the 2014 formal test. Table 18.7 also shows that the 20 students' average total scores on the 2009–2013 TEM-8 papers did not vary much (6.2–8.87 points) from their scores on the 2014 formal TEM-8 test except for the scores for the 2009 paper, which is 14.80 points lower than their scores on the 2014 test.

Based on the above research results, this study proposes three suggestions for the current TEM-8 test paper design:

(1) In light of the rising influence of the TEM-8 over the past two decades as a large-scale, high-risk, and the highest level language comprehensive ability test for English Majors in China, more attention should be paid to the stability of its difficulty level, particularly that of the subjective items, i.e., translation, proof-reading and error-correction, and writing, to ensure its fairness. These four parts account for half the full score of the test and should maintain a certain level of stability in difficulty.

(2) According to the 15 hypotheses proposed by Alderson and Wall (1993) for producing washback on teachers' teaching and students' learning, the "PPP" washback model proposed by Hughes (1993) and the implicit and explicit backwash forms proposed by Prodromou (1995), it is here suggested that the answering form, the questions design and options for objective questions be further improved, and more test items related to the specialty characteristics of the English major and application capacity should be included so as to promote teaching and stimulate students to form good learning habits, develop learning strategies, and improve their comprehensive language ability.

(3) To ensure the reliability of TEM-8 questions, more attention should be paid to such common practices as pre-test conduction, test item analysis, and post-test work covering testing scoring and equating under the guidance of scientific language testing theory implemented using Rasch-based on and IRT–based software (Zhang 2013). Only in this way can the quality of TEM-8 test improved to a greater extent. By then, the TEM-8 certificate is expected to be more recognized not only in China, but also in the world as well.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 4*(2), 115–129.

Bachman, L. F., & Palmer, A. (1996). *Testing in practice*. Oxford: Oxford University Press.

Bao, X. (2009). A study of the validity of interpretation test under the communicative language testing theory-a cast study of shanghai advanced and intermediary interpretation posts qualifications certificate exam. *Foreign Languages World*, (4), 84–90.

Brown, J. D. (2006). *Foreign language teacher education and development series-experiencing English teaching series: Testing and evaluation in language project*. Beijing: Higher Education Press.

Gui, S. (1986). *Standardized test: theory, principles and methods* .Guangzhou: Guangdong Higher Education Press.

Guo Li. (2010). A study of banked cloze test. *Foreign Languages in China*, (4): 70–76 [5].

He, L., & Zhang, J. (2008). A study of the reliability of CET band-4and band-6 oral exams under the multi-dimension Rasch model (CET-SET). *Modern Foreign Languages*, (4), 387–398.

He, Y. (2005). Validity of the construction of listening tests and its realization. *Foreign Language Teaching*, (3), 72–74.

Hughes, A. (1993). *Backwash and TOEFL 2000 unpublished manuscript*. UK: University of Reading.

Li, H. (2011). A study of the reliability of the CET-6 writing scoring under the generalizability theory and the multi-dimension Rasch Model. *Foreign Languages and Foreign Language Teaching*, (5), 51–56.

Li, Q, & Kong, W. (2009). Computer-based language testing and its validation verification. *Foreign Languages World*, (3), 66–96.

Liu, J. (2010). A study of the multi-dimension Rasch Model of the test paper rater effect. *Modern Foreign Languages*, (2), 185–193.

Prodromou, L. (1995). The backwash effect: from testing to teaching. *ELT Journal, 49*(1), 13–25.

Qin X. (2012). Making full use of the washback effect of the TEM-4 and TEM-8 to prevent the decline of teaching quality. *Foreign Language World*, (3), 10–14 + 41.

Wang, H. (2007). A study of the validity of the oral test of TEM-8 based on scoring process evidence. *Journal of PLA Foreign Languages Institute*, (4), 49–53.

Wen, Q., & Wang, L. (2009). The validity of the oral test of TEM-8. *Journal of PLA Foreign Languages Institute*, (5), 7–41.

Xiao, W. (2011). A study of the rater reliability—A survey of the TEM-8 translation raters. *Foreign Language Journal*, (6), 115–119.

Xu, Q, & Zhang, Y. (2004). An analysis of the content of the TEM-8 grammar and vocabulary part. *Foreign Language Research*, (2), 57–59.

Yang, W. (2011). Validation of the answers to reading questions in the TEM-8 tests. *Foreign Language Teaching*, (6), 53–56.

Zhang, Q., & Yang, H. (eds.), (2013). *Pacific Rim objective measurement symposium(Proms) conference proceedings.*Berlin Heidelberg: Springer.

Zhang, S., & Yu, P. (2010). A study of the reliability guarantee of the reliability online marking of TEM-8 writing. *Foreign Languages World*, (5), 79–86.

Zou, S., Zhang, Y., & Zhou, Y. (2002). The relationship between the reading test question type, strategy and scores: a study of the validity of the answers to TEM-8 reading questions. *Foreign Language and Foreign Languages Teaching*, (5), 19–22.

Chapter 19
The Application of Descriptive Statistical Analysis of Foreign Language Teaching and Testing

Wu Wei and Li Lan

Abstract In the traditional way of language testing, teachers usually analyze the scores by calculating the mean and classifying point segments. Though easy to operate, the method of mean calculation is susceptible to extreme data and the classification of point segments cannot tell the distribution of scores when they are abnormally distributed, resulting in the unavailability of positive feedback on teaching from the test. Therefore, this paper proposes the use of Descriptive Statistical Analysis and the Combination of Measures of Central Tendency, Divergence Tendency, and Standard Score, which will find out valuable information to enhance the teaching quality.

Keywords Descriptive statistical analysis · Central tendency · Divergence tendency · Standard score · Foreign language teaching

19.1 Introduction

With the development of language teaching, language testing is increasingly becoming a hot research area. In its continuous development process, how to apply a language test to measure language ability has been a core issue. In the field of foreign language teaching and research, language testing was considered the most effective means of measuring language proficiency, and at the same time, it produced the most data. Therefore, linked language testing and statistical methods together and the application of statistical methods in the field of language testing are inevitable. Only by using the scientific statistical methods can we make full, true, accurate, and reliable analysis of the data, and to provide analyzed materials for improving the quality of language teaching and measuring language ability.

There are all kinds of test in foreign language teaching, for teaching is indispensable to test and the subsequent calculation of results and data analysis. Some

W. Wei (✉) · L. Lan
Hebei University of Economics and Business, Shijiazhuang, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_19

famous universities abroad, GRE and TOEFL and so on have employed Standard Score to assess the examinee scores, while the assessment method is relatively lag behind in domestic. In the examination of the past, the analysis methods of performance teacher used are single and the content is ambiguous. Direct at the statistical analysis of results in each test, many foreign language teachers mainly calculate two central tendency data–the mean and point segments, as to standard deviation and standard score, they know little about that. This is not only goes against the exploring and feedback of teaching information, but also affects the teaching decision making. Toward the defects of the traditional analysis method, this paper introduces a simple and practical descriptive statistical analysis on the basis of the research experiences of overseas study. Through scientific, rigorous, quantitative analysis of the case data, we found out distinct consequences from traditional method on evaluating student performance, which truly and objectively reflect the students exam results. The case data come from three English final exam scores of grade 06 undergraduate class in Hunan University, we extract some representative scores as samples.

19.2 Theoretical Basis for Applying Statistical Methods in the Field of Foreign Language Testing

With the rapid development and improvement of foreign language testing theory and methods, intersection with psychological measurement theory has become a commonplace. Roughly speaking, language testing has gone through four distinct as well as overlapping stages, i.e., the pre-scientific stage, the psychometric structuralist stage, and the psycho-linguistic socio-linguistic stage (Spolsky 1999) and the communicative pragmatic stage (Bachman 1990). The statistical methods began to be used in the psychometric structuralist stage. Bachman (2004) divides communicative language ability into language competence, strategic competence, and psychological mechanism in the communicative pragmatic stage, and highlights the importance of mental capacity to language testing and research. It then became an important part of language testing, and the three cornerstones of psychological measurement theory provided a theoretical basis for statistical methods in the field of foreign language testing, thereby promoting statistical methods plays a growing role in the field of foreign language testing.

19.3 The Application of Descriptive Statistical Analysis of Foreign Language Teaching and Testing

Traditional analysis method of students' performance mainly adopts central tendency—the content is relatively single which cannot fully reflect the problem. In pursuit of the scientific of measurement, descriptive statistical method includes two more evaluation method—divergence tendency and standard score on the basis of the original analysis. It can explore the results' information profoundly and reflect problems in a more comprehensive and objectively way.

19.3.1 Central Tendency

Central tendency means the statistics of data toward the center place, which reflects the concentration of data. Central tendency is the representative value of a set of data, which can be used to illustrate a certain characteristics of a whole set of data, namely their typical case. Central tendency statistics mainly includes mode, median, and mean. Mode refers to the most frequently value in a variable, can indicate the central tendency of nominal level variables better. If you put the values of a variable from low to high in sequence, the one in the middle called median. The median divides the data into two equal values, which means there are 50 % observed value above it and the other 50 % below it. The median can reflect the level mathematical characteristics of ordinal variables well, but not easy to response to the change of the numerical. Average refers to the arithmetic average of a set of variable data. The most commonly method of computing averages we used is average value *X*, for its formula

$$\overline{X} = \frac{x_1 + x_2 + x_i + \cdots + x_n}{n}$$

'*Xi*' represents the performance of classmate number *i*, '*n*' is the total number of the class. If the statistical grouping of the whole class's performance has already been made, then takes the formula:

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + x_i f_i + \cdots + x_n f_n}{f_1 + f_2 + f_i + \cdots + f_n}$$

Among them: *xi* represents the values of group number *i*, '*fi*' stands for the number of group *i*.

Each statistic of central tendency has its advantages and disadvantages. Unlike the median only reflect the middle piece of the data distribution, the calculation of average involves every data. Mode is not always in the middle, so it may not be representative. Overall, average is the most reliable, the most representative and the most widely applied central tendency. Besides that, its calculation is simple and easy to operate. However, average is also easily affected by extreme data, and the representative will weaken if there is extreme value. The remedial method we can adopt is trimmed mean rule out the extreme value of both ends in proportion. In addition, if the class-interval is uncertainty in grouped frequency (frequency distribution), average will not be able to work out. Mode and median are not subject to this restriction. In general test statistics, we can use average combined with median

Table 19.1 The contrast of mean, median and mode

Exam scores						Mean	Median	Mode
Group A	75	80	90	95	101			
	108	110		115	117	107.92	110	117
	117	124		131	140			
Group B	70	88	94	94	95			
	110	116		117	119	107.92	116	94
	121	124		126	129			
Group C	86	95	96	98	100			
	100	105		106	100	107.92	105	100
	115	124		129	139			

and mode to exert the characteristics of various statistical, which fully reflects all kinds of data we want to understand and provide much more detailed information for test result analysis. Typical cases are different during the mean, median, and mode. As shown in Table 19.1.

19.3.2 Divergence Tendency

In order to avoid the influence of extreme data, we should also report the discrete quantity of the data, namely divergence tendency. Looking at the same grade of several classes, for example, the average scores of a few classes are the same or very similar, but may be different on the score distribution. It cannot comprehensive and truly reflect the full picture of learning about college English course if only compare the average. Only measure the discrete degree of each class's achievement scores at the same time, can be more comprehensive describe and reflect these classes' actual situation of. Therefore, when we study the characteristics of a set of data, not only understand the typical case, but also understand their special circumstances. Only in this way, can we understand the differences between data much clearer.

19.3.2.1 Range

So-called Range refers to the difference between maximum and minimum in a set of data. Range can let us know the degree of the divergence. The larger the Range is, the higher the degree of discrete data. In turn, the representation of central tendency statistics became smaller. Since range only depends on the size of two extreme values and ignoring the other numerical distribution, therefore, with range only roughly estimate the discrete tendency. As a result, people tend to use standard deviation to test the discrete situation of data.

19.3.2.2 Standard Deviation

Standard deviation is also called the mean square error, which is the square root of variance. At present, Standard deviation is the most commonly applied the most scientific index for measuring discrete degree of results. Variance is defined as the mean after the square of deviation from average. That is, the mean value after square of the difference between each data with the average number of the group. Standard deviation "is the best indicator to represent the difference degree of exam results, the larger its value, the greater the discrete degree of the fraction distribution; the smaller the value, the smaller the discrete degree and the concentration of fraction distribution".

$$SD = \sqrt{\frac{\sum (x_i - \overline{X})^2}{n}}$$

Too big or too small of standard deviation suggest that there exists abnormal condition in the exam. A large number of statistical practices show that when examination results are normally distributed as shown in Fig. 19.1: range is about six standard deviations. But under normal circumstances, range is commonly about half of full marks. That is to say, if full score is 100 points, scores of discrete degree is more reasonable when standard deviation is 10–15 points.

In the application of central tendency when describe the typical case of a set of data, there is the problem about the size of representation of central tendency, as shown in Table 19.2.

Three sets of results in the table are the average of 107.92 points, but the range in group C is 53, its standard deviation is 15.135, very concentrated; Group B in the middle, its range is 59 and the standard deviation is 17.947. Range of data in group A is 65, the standard deviation is 19.233, most scattered; in other words, the representation of group C is maximum, group A is minimal. If only judging from the average of three groups, there are no comprehensive conclusions. Thus, the representativeness of central tendency in a set of data can be illustrated by the measure which represents the difference, such as the range and standard deviation in the table. The smaller the dispersion tendency is, the greater the representative of

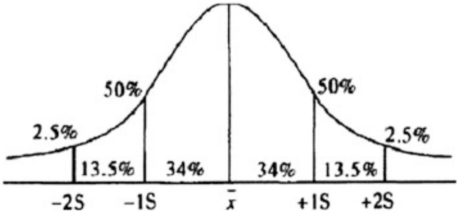


Fig. 19.1 Bachman typical normal distribution of norms referenced test

Table 19.2 The contrast of mean, range and standard deviation

Exam scores						Mean	Range	Standard deviation
Group A	75	80	90	95	101			
	108	110		115	117	107.92	65	19.233
	117	124		131	140			
Group B	70	88	94	94	95			
	110	116		117	119	107.92	59	17.947
	121	124		126	129			
Group C	86	95	96	98	100			
	100	105		106	100	107.92	53	15.135
	115	124		129	139			

central tendency, also the greater the concentration of data; on the contrary, if the dispersion is larger, the smaller the representative of central tendency and the greater the degree of discrete data. If the dispersion is zero, that means the group data equal to each other and their values are equivalent to central tendency.

19.3.2.3 Standard Score

Now in school routine examination, centennial system is the most widespread employed to evaluate test scores objectively. However, due to the different difficulty, different classes, different time of examinations, raw score is not comparable, so the limitations of centennial system are great. In order to realize the real comparison between academic achievement and learning ability of examinees, we must translate the raw score into comparable standard score and obtain the information that in favor of teaching management and decision making. Standard score (also called Z point) is actually represented the difference between a certain score and the average in standard deviation units. The calculation formula of standard score is:

$$Z = \frac{x - \overline{X}}{SD}$$

Standard score can not only show the status of each student's performance in the entire distribution, but also compare with raw data in different distribution. Now we illustrate the application of standard score with specific example. The final exam scores of student A and B, as shown in Table 19.3

Look from the raw scores, B's result is higher than A, it seems that B is better on performance. But the total points of A's standard score is 2.25, and B is −0.5, which signifies that A's relative level is far higher than B. From the above example, you can see there often with a decimal after the raw scores translate into standard score.

Table 19.3 The contrast of raw score and standard score

Items	Raw score		The whole class		Standard deviation	
	A	B	$\bar{X}$	*SD*	A	B
Listening	78	72	77	2	0.5	−2.5
Speaking	87	80	80	2	3.5	0
Reading	70	89	81	4	−2.75	2
Writing	68	65	65	3	1	0
Total score	303	306			2.25	−0.5

Moreover students with results lower than the average, their standard score is negative. To avoid this situation, people convert Z scores to further T scores, that is, the average of standard score is no longer set as 0, but 50; the standard deviation of standard score is no longer as 1, but 10. Expressed in formula:

$$T_{\text{score}} = SD * Z + \overline{X} = 10Z + 50$$

In above example, A's standard score is 2.25 after converted to T score. A is 72.5 points and B is 45 points, this result is contrast to the traditional average value. So we should not view things from the surface phenomenon, but reflect students' actual level in a more scientific and sample plot way.

19.4 Future Trend of Applying Statistical Methods in the Field of Foreign Language Testing

From the above data and analysis, on the one hand, for small-scale tests, in the case which is not very accurate, the use of descriptive statistics is very practical, relatively simple and easy to operate. As a traditional method, it has developed relatively complete. However, the estimation of item difficulty, discrimination and reliability depends on the testees group, and the estimated capacity of the testees depends on the test items, it can only provide average measurement accuracy, if all the testees passed or failed, the indicators such as facility value and item discrimination would lose its meaning. On the other hand, in the field of foreign language testing, more and more scholars tend to accept multi-dimensional view of language which considers language ability has multiple dimensions, and that different language skills are a combination of different cognitive abilities. Therefore, the demand for the application of inferential statistical methods which can estimate the relationship and interactions among various factors were growing rapidly. It is visible that in a certain period of time, descriptive statistics and inferential statistics will complement each other and develop together, the theory and application of inferential statistics will also tend to improve gradually, and the dependence and cooperation with descriptive statistics will also increase. In addition, according to the data and the current research status in the field of foreign language testing, in relation to statistical methods, the application of statistical software has become a trend. Currently, the important foreign language tests usually implements the way of intensive testing which has the characteristics of more testees, miscellaneous exam rooms, a wide range of test scope and so on. After each test, a large paper handling capacity was needed, related to classification, marking, scoring, registration points, statistics, analysis, and a series of operations which are very cumbersome. If the collation and calculation of data is by hand, the work is both arduous and tedious, and this can only provide the basic information such as the test scores, the mean, the pass rate, and so on. This will lead to a series of problems such as the data processing method being too simple, inadequate comparing parameters, arguments, waste of information resources and the backwash effect of test cannot fully play its role. In order to compare the results of tests, make the analysis of tests digitized, accuracy, scientific, and efficient, let teachers get acquaintance with his teaching quality and try to improve the teaching, we must apply computers to manage and analyze the test scores of testees. The computer is extremely versatile, so it is suitable for foreign language testing in any size, any groups, and any courses. As computer has friendly interface, simple operation system, the role of it in the field of foreign language testing is bound to become increasingly significant. At present, the computer adaptive test, the TOEFL computer exam, etc. are all good examples. In the related sample articles, when making statistical analysis, many of them directly or indirectly applied related statistical software. This has provided many conveniences for the research and improved efficiency, saved time, and reduced the occurrence of mistakes. It is obvious that with the development of testing research and the statistical theory, the shortcomings and problems of traditional statistical methods have been overcome and the testing accuracy is improved. At the same time, the process of statistical analysis has become increasingly complex; the conditions for implementing statistical calculation also increased, and in most cases it has to rely on computer software and procedures to complete a variety of complex statistical and mathematical calculations. In actual teaching process, with the assistance of computer, language teachers can select the appropriate and feasible methods according to test characteristics, test objectives and the practicality of test, so that the test can better serve the teaching. In the wave of applying statistical methods in the field of foreign language testing, the application of computers and statistical software has already become very common. With the development of the test theory and the popularity of quantitative research, the unique aspects of advantages of computers in handling the test data will be further demonstrated, and the dependence on the computer of language testing scholars will be growing.

19.5 How to Make Good Use of Statistical Methods in the Field of Foreign Language Testing

Although there are many problems that call attention in the application of statistical methods in the field of foreign language testing, there are two basic issues needing extra attention. (1) For descriptive statistical methods, it is important to distinguish the relationship between percentage grading system and standard deviation score. In the same test, the raw score can be used to judge the testee's language ability, but in the different tests, the raw score should be transformed into standard deviation score. The introduction of standard deviation score can eliminate item difficulty's effect in the total score on the weight of different tests, reflect the weight of different tests in the total score, and ensure the role of different tests in the total score. Standard deviation score takes the standard deviation as the basic unit to measure the deviation between every raw score and the standard deviation, no matter how different of the mean and standard deviation of different tests. After transforming into standard deviation score, they will become the fixed form with the same unit as the basis, so the limitation of raw score will be overcome and the calculation and comparison of testee's language ability will be more rational. In the aspects such as the degree of association with raw score, the stability of information, and the fluctuation of scores are in transformation, this kind of linear transformation is significantly better than other conversion technology, and therefore it is more conducive to the interpretation and use of scores, and to the selection of talents. It preserves the merits of the absolute scores and relative scores, especially, the relative scores, and its scientificity and rationality have gained more and more people's attention. (2) For inferential statistical methods, the special attention should be paid to the sample requirements, or it will make no sense. Since the inferential statistical methods are usually applied based on samples, the hypothesis of normal distribution of samples is very important. Therefore, for inferential statistics, to ensure the reliability of statistical analysis result, the normality test is required, and if the data was not normally distributed, normal conversion has to be delivered first, and after that, to conduct a variety of statistical tests, to make the relevant inference, otherwise, the results are meaningless. In addition, different statistical techniques need different samples and the sample size should also be focused on. What's more, as scholars in the field of foreign language testing, we need to acquire a certain amount of statistical knowledge and establish a scientific spirit, and thus to better and correctly apply statistical methods and computers for the service of foreign language learning, teaching, and research. First, the enhancement and improvement of the foreign language practice abilities should be placed in the first place, it is necessary to constantly face the new situation, new challenge sand adapt to change. Because of the characteristics of test target and testees, the design, development, analysis, assessment, and the result report of the test tools (examination papers) have to rely on certain statistical theory. Thus, the test is coated with a layer of mystery, making the foreign language teachers and researchers prohibitive to the general scientific testing theory, which led to on the one hand, teachers completed the proposition of test paper on its own experience and cannot guarantee the quality of the paper, and this is unfair to students either, on the other hand, due to lack of necessary statistical knowledge, it makes the test results in waste of a large amount of hidden information, or improperly interpreted the test results. Therefore, to ensure the correct application of statistical method in the field of foreign language testing, it requires the scholars to learn and master the basic knowledge, basic operations, foster a scientific spirit on the application of statistical methods. Knowledge of psychology and pedagogy should be learned and psychological law, study law should also be mastered to teach students in accordance with their aptitude. Modern teaching assistant tools and methods like computers need to be applied in teaching and researching. Although for the scholars in foreign language testing, there are different levels of requirements, the basic requirement is to master the basic theory, basic concepts, basic rules of language testing, and how to design the test items, what problems should be paid attention to during the designing. If you want to deepen the research in language testing, then you need to be able to master some basic statistical knowledge for language testing. Only by a comprehensive analysis of the test results, a scientific research on test items and the teas paper, the problems of teaching, learning, and test paper designing can be found. And this can make future foreign language testing more rational and scientific, so it can play its role in promotion of teaching and learning.

19.6 Conclusion

It goes without saying that the descriptive statistical analysis method is more scientific and reasonable—for it overcomes the disadvantages of the traditional statistical analysis and provides us with rich teaching information. In addition, it also provides a theoretical basis for seizing the examination quality on the whole: on one hand, it contributes to the dynamic management of the quality of teaching; on the other hand, it is beneficial to promote students' learning consciousness. With the development of the computer testing technology, this kind of descriptive statistical analysis will be more and more popular. At the same time we should also note that descriptive statistical techniques used in the test questions and examination paper analysis remains to be supplement and perfection in the content, it will only make us to know part facts, some problems still need make further inferential statistics method, such as chi-square test or variance analysis. In a word, to truly realize evaluation of test questions and examination paper scientific and objective, there is still a long way to go.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing [M]*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment [M]*. Cambridge: Cambridge University Press.

Spolsky, B. (1999). *Measured Words*. Shanghai Foreign Language and Education Press.

Chapter 20
Backwash Effect of English Testing on English Teaching for Adults

Yuhuan Gong

Abstract At present, there have been a great many researches on backwash effect of English testing on English teaching in Universities, few of which involves English teaching in Adult colleges. This paper discusses the quality criteria of the English testing like the validity, the reliability and so forth, referencing to the research and the theory on the relationship between language teaching and testing at home and abroad. The paper also reveals the backwash effect of English testing on English teaching in the adult college. Meanwhile, reform proposals to English testing have been made in order to bring the positive function of English testing into play and to improve English language testing scientifically.

Keywords Test · Backwash effect · English teaching for adults

20.1 Current Situation of Public English Teaching and Testing in Adult Colleges

20.1.1 Foreign Language Teaching and Testing

There is a close connection between foreign language teaching and testing. Foreign language testing, as an important factor in the process of foreign language teaching, can be used to evaluate the quality of language teaching, learn the situations of student learning, and adjust and improve subsequent stages of foreign language teaching according to the test results. Zou Shen divided testing functions into teaching and research, and teaching is further subdivided into the backwash effect, academic examination function, selection function, and evaluation function. Backwash effect means that language testing is the way to test teaching effect and teaching objectives, at the same time in turns affects the formulating and adjusting

Y. Gong (✉)
English Department, Beijing Dongcheng Vocational University, Beijing, China
e-mail: gongyuh2004@163.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_20

Table 20.1 13 level of public English test questions

Category	Type	Score	Note
Listening (35 %)	Listen to the sentence and response	10	Objective question
	Dialogue	20	Objective question
	Short passage	5	Objective question
Grammar and vocabulary (30 %)	Grammar	25	Objective question
	Vocabulary	5	Objective question
Reading comprehension (40 %)	Reading	35	Objective question
	Task reading	5	Subjective question
Translation (15 %)	Translate English into Chinese	5	Subjective question
	Translate Chinese into English	5	Subjective question

of the teaching goals and contents. In the process of foreign language teaching, how to maximize the positive testing backwash effect, restrain its negative effects, has become a common concern of foreign language teaching researchers.

20.1.2 The Characteristics and Requirements of Public English Teaching in Adult Colleges

Public English teaching in adult colleges has its own characteristic and the request. In the teaching goal, public English teaching pays more attention to the cultivation of learning and practical ability, sustainable development ability of the students. English teaching in adult colleges emphasizes language function and application of language. Training basic skills of language and application ability in communicative activity are both important. However, in the actual teaching process, the public English teaching in adult colleges is faced with many problems. The author thought that the biggest problem is—Basic English teaching is overemphasized, and English for specific purposes (ESP) teaching is obviously lagging. Limited to the teachers, teaching methods, teaching material development, most of the current public English Teaching in adult colleges is to teach the basic language knowledge and skills. ESP teachers are scarce. Some contents of ESP teaching are too hard for students to learn.

20.1.3 The Situation of Public English Testing in Adult Colleges

Evaluating the quality of a specific test, we should start from its validity, reliability, backwash effect, and other index. In addition to the above standard, we should know whether the distribution of testing items in the whole testing system is reasonable. And from a macro perspective, we should examine whether the whole test system is consistent with the final teaching objectives. The following is the data analysis of public English test in my school (Tables 20.1, 20.2 and 20.3).

The author thinks public English testing in adult colleges mainly has the following problems:

Table 20.2 Public English test statistics—in 2014 January test as an example

The number of participants	Average score	The excellent rate	The pass rate	The low rate
256	70.99	12.7 %	56 %	12 %

Notes 95–100 is excellent, above 600 is pass, below 65 is low score

Table 20.3 Analysis of public English test content

Category	Type	Score	Testing goal	Proportion (%)	Interactive
Listening (35 %)	Listen to the sentence and response	10	Skill of listening to sentence	8.33	No
	Dialogue	20	Skill of listening to dialogue	16.66	No
	Short passage	5	Skill of listening to short passage	4.17	No
Grammar and vocabulary (30 %)	Grammar	25	Language knowledge	20.83	No
	Vocabulary	5	Language knowledge	4.17	No
Reading comprehension (40 %)	Reading	35	Reading skill	29.17	No
	Task reading	5	Reading skill	4.17	No
Translation (15 %)	Translate English into Chinese	5	Writing skill	4.17	No
	Translate Chinese into English	5	Writing skill	8.33	No

Notes Here the "interactive" refers to the communication in real or close to the real language context, are not the "interactive" between test-takers and language materials

(1) From the testing structure, test attaches too much importance on the assessment of ordinary language knowledge, lacking of the assessment of English application ability. The language used in testing is different from the language in real life and cannot reflect the authenticity and practicality of English language.
(2) Common English and ESP test cannot be properly integrated, and testing content is lack of "professional". To carry out occupation oriented tenet in adult colleges, teaching is required to maximally meet the job requirements. So in adult colleges, the curriculum system of public English teaching should be based on the work process, and establish a set of evaluation system. However, at present the public English testing generally ignored the assessment of ESP, and cannot integrate the content of common English testing and ESP testing successfully.

20.2 Build "1+X" Testing System

20.2.1 Construct the Principles the Public English Testing System Should Follow in Adult Colleges

Reforming the traditional test mode and trying to set up the new testing system, is to improve the testing quality, make the backwash of testing play a positive role. According to the latest research on the foreign language testing theory, Hughes put forward some suggestions to increase testing positive backwash effect, worthy of our reference.

(1) Test skills which are hoping to obtain the development;
(2) Try to use the direct measuring method;
(3) Take the ability standards as the basis of testing;
(4) The examination results should be combined with the teaching objectives;
(5) Training and guidance to teachers when necessary;
(6) Taking into account the examination fee.

We believe that, in the construction of public English testing system in adult colleges, it is necessary to consider the above elements. On one hand, we should always adhere to the combination of the test itself and public English teaching objectives. Testing content should examine students' English practical ability. On the other hand, we should make the test mode diversified, use direct test method as much as possible. In this test, students will be required to directly use the skill or ability, such as oral, writing, translating ability. This requires that it may be necessary to separate test of spoken English, writing to evaluate productive skills. In addition, we should also take into consideration the test cost problem. The population to take part in public English test is very large. I am afraid it is difficult to put the test design and conception into practice if we lack of manpower, material resources, financial support. So in the building of public English testing system in adult colleges, we should simplify the procedures and steps of the test at same time ensure the quality of test.

Finally, we should also pay attention to the training and guidance of the examiners, especially to the test proposition work. Generally speaking, English mid-term test and the end-of-term test proposition work were done by the English teachers themselves. Strict requirements on the proposition and standard will be effective to ensure the quality of Public English test.

20.2.2 The Construction of "1+X" Testing System

A complete set of testing system is composed of concrete test organic parts. These specific test items follow certain sequence according to a certain proportion. The final purpose of the test as the main line, the test organic parts constitute a complex and integrated test system. In adult colleges, public English test system should include testing purpose, contents, and test methods and operation.

(1) The ultimate goal of the design of the testing system should examine the students' application of English. Combined with the particularity of the occupational education, testing should be associated with the professional English. Public English teaching should include the English used in the process of work. The author thinks workplace English test should become an important part of Public English testing system, but the testing principle should be formulated and improved further.

(2) In the public English test system, the test method is a vital issue. Selecting what kind of examination method, using a single test mode, or multiple tests, these problems we need to carefully consider and treat. We need to take into account three aspects that Zou Shen talks, the reliability and the validity of examinations, and the operation of examination. In order to realize the balance between the three aspects, we should take the direct test and indirect test when we designed an English test system. The direct test is used to improve the test validity and the indirect test is used to improve the reliability of the examination.

With the above ideas, this paper proposed the "1+X" test system of public English in adult colleges. The core of the system is "1" and "X". Here "1" refers to the basic knowledge and skills of English which students should master. The "1" is to measure the basic language ability of students. Its main purpose is to check the students' language ability in a time of public English process (such as a semester or academic year after). The "1" assessment is similar to Proficiency Test. Propositional principle, type design, the structure of the test can take PRETCO as a reference. The author thinks that adult colleges can develop a public English test database.

The test database is required to ensure to have various types, adequate testing items, regular adjustment and change one of the testing items. In the design of a specific set of paper, we just select testing items from testing database according to certain principles and methods and assemble them into a complete set of papers.

If the “1” is seemed as the internal quantification in public English test system, the “X” is the variables in the test system. In adult colleges, the evaluation of the students’ occupational English should be included in the public English test system. Besides the basic language knowledge and skills assessment, we should consider ESP assessment, this is what we call “X”. Of course for the students of different majors, the content of their professional English assessment should be different. According to different majors, students can be divided into several major English testing groups. Such as group from Preschool Education, Hotel Management, Accounting, Finance, and so on. The “X” here mainly refers to groups from different majors.

The requirements of common English provide guarantee for the sustainable development of students in the future, and the requirements of ESP are helpful to students’ occupational ability. What kinds of ways to evaluate the students English occupational ability? The author thought different colleges can adopt different flexible way. Of course, to test the students’ English professional ability, we should pay attention to the following points:

(1) The assessment of students’ professional English should be mainly the listening and speaking. If the test conditions allow, it may be the oral test. The test result is an important index for evaluation of the students’ English learning results. Liu Runqing, Han Baocheng pointed out that a person’s oral English is very important from the point view of social demand for talents. The testing is not comprehensive if it has no oral language testing or it cannot be called language testing.

(2) The assessment of students’ professional English can take 100 % formative assessment. At various stages in one semester (or a year), homework which could be the form of task can be assigned to students as the formative assessment. Students in group finish the homework and show their homework in the form of report or role-play in the class. The formative assessment focuses on the professional English skills and oral English ability.

20.2.3 Some Suggestions on Carrying Out the “1+X” Test System

“1+X” test system in adult colleges not only guarantees the evaluation on students’ overall English level, but also check the students’ professional English ability. Refer to the implementation of the test, the author thinks that we need to pay attention to the following points:

(1) We should guarantee the quality of the assessment which is on students' overall English. This requires us to establish a test specification, detailed provisions on what and how to test. Here the author suggests a public English test proposition team can be established in adult colleges. According to the test purpose, test content, test methods and other relevant factors, the proposition team should make the proposition specific which can guarantee the uniform of each set of paper in a certain extent.

(2) We would better to take oral examination to evaluate the students' professional English. Objectively speaking, compared with several other language skills test, oral exam questions are difficult to operate. It is not because oral exam is difficult to design, it is because oral English test is time-consuming and laborious and oral English test score is also very difficult to be completely objective and fair, with greater subjectivity. Therefore, before the oral English test, it is necessary for the examiner to have training and guidance, to clear the testing step, process, form, scoring method, and standard, etc. Here in particular the scoring method of oral English test mainly has the holistic scoring and analytic scoring. The author thinks that in the process of test, we can combine the two methods, and establish a unified score list. In addition, according to different students group, part of the test content should also be differentiated. This requires the proposition personnel notices differences in the design of oral test questions for different professional group and make many different sets of test more uniform and parallel.

(3) The evaluation on student's English overall qualities and professional English is an important part of the whole public English test system and an important index for judging the student's public English learning results. In the final results of an examinee public English test, the test on common English and on the professional English should be combined according to a certain proportion. This ensures that the integrity "1+X" test system, also the consistency and balance of the "1+X".

20.3 Conclusion

Foreign language testing has many connections with foreign language teaching. Testing and teaching promotes and restricts each other, especially test has the backwash effect (positive or negative) on teaching. In china, foreign language testing has great effect on foreign language teaching. So how to make the foreign language test promote foreign language teaching and become the necessary part of foreign language teaching is a meaningful topic. As a necessary course in adult education, public English teaching has the important mission of cultivating students' English practical ability. However, the current public English test mode in adult colleges is difficult to play its positive backwash effect. So it is necessary to reform the public English test system, build a reasonable and effective test system.

In view of this, this paper puts forward the “1+X” public English test system. This test system evaluates the students’ common English quality and professional English quality and put emphasis on listening and speaking ability. We think this testing system has wide applicability and maneuverability. Of course, the test system requires public English teaching to adjust and adapt to the test. In all, teaching and testing can promote each other.

Chapter 21
Recasting English for Environmental Science into a SPOC—A Pilot Study in Wenzhou University

Xiaoshu Xu and Xiuyu Gu

Abstract With the economic globalization and the internationalization of higher education, a new demand for social development and a change of needs of disciplinary development were initiated which require the reposition of College English in China. This paper suggests that College English in China should change toward English for Specific Purpose which intends to cultivate learners' professional or academic English communication skills. In Wenzhou University, ESP has been carried out since 2009. However, a survey for graduates showed a giant gap between ESP program requirements and students achievement. To improve the efficiency of ESP teaching, the research group, based on Knowles' Andragogy Theory, starts to recast English for Environmental Science into a "SPOC+Tradition Education" format. The purpose of the experiment is to improve students' performance and work-related abilities through active learning. The approach has two steps. The first step is to construct a SPOC platform and remaster the course with the help of transdisciplinary professors, industry experts, and information technology experts. The second step is to spread the new concept of SPOC for the on-site course learning among the lecturers, administrators and students as well as to undertake practical training for them. Altogether, 119 students of first-year bachelor majored in Environmental Science in Wenzhou University have taken part in this pilot study. The result of the questionnaire survey shows that the majority of the subjects is satisfied with the remastered course teaching, and believes that the new format of learning can inspire their learning interests, improve their learning strategy, and help them adapt to collaborative learning model.

Keywords English for environmental science · SPOC · College English reform · Knowles' Andragogy Theory

X. Xu (✉)
Wenzhou University, City University of Macau, Macau, China

X. Gu
College of Education, City University of Macau, Macau, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7_21

21.1 Introduction

Since the end of twentieth century, China has embarked on a policy of rapid expansion in higher education, and the concept of “University of Applied Science” was put forward with the adjustment of industry structure national wide (Tan 2013). The development of “University of Applied Science” was in line with the national strategy to create a better trained workforce. In February 2014, Premier Li Keqiang at the State Council requires to guide a group of local universities to transform into university of applied science in the State Council executive meeting. The reform policy for local Universities and the fast-changing world challenge the University to reconsider new forms of learning, new technologies for teaching and new requirements for graduate competence. Educators are challenged to design new learning environments and curriculum that can really encourage students’ motivation and facilitate students’ active learning. More important, educators need to ask if the skills imparted are really transferable to the workplace (Tan 2013).

The first wave of College English reform in China started decade ago. Until now, the English proficiency of college students has generally improved. In this situation, General English Course designed for improving students basic language skills is no longer necessary. Zhao have carried out a survey on students English learning interests among 2283 freshmen and sophomores in 12 universities, the result shows that 34.8 % students have no interest in English learning. A survey made by Yu and Yuan shows that the major obstacles students met in college English learning are lack of goals, interest, and pressure in learning. General English Course can no longer satisfy different majors and employers’ requirements on using English to do academic research and work. The British Culture Committee had carried out a wide range survey named “English 2000” in the end of last century, the result of which shows that nearly 90 % experts believe that English teaching in the twenty-first century should be combined with other disciplines rather than English as a language. The distinguish feature of language learning is leaning by doing, thus, using English to deliver knowledge of other discipline is the best choice.

Universities should scientifically determine the target of English teaching according to the standards of talents cultivating in each university and the actual needs of students’ future jobs. As for the local universities, the English teaching objectives should be English proficiency+expertise=English capabilities to attend professional activities and international exchanges. According to this goal, for the majority of local Universities, college English teaching should be to train students in reading and writing, while special attention should be paid to students’ ability to use English in professional fields. Only by combining English skills and expertise in College English teaching, will it stimulate learners’ interest and motivation to improve learning efficiency. Wenzhou University has chosen ESP since 2010. ESP is the continuation or expansion of basis English teaching, which aims to further develop students’ practical language abilities based on general English teaching when students language knowledge and skills have developed to a certain stage. ESP teaching objectives and teaching content are determined by learner's pragmatic language capabilities or actual use of English rather than general education (English as a discipline) (Strevens 1977). ESP aims to train students to use English when carry out their professional work, which meets the needs for the compound talents under the economic globalization background.

A survey (2013) for graduates carried out in Wenzhou University indicates that a good number of graduates declare that their English are incompetent for their profession. To improve the efficiency of ESP teaching, it is suggested that Knowles' Andragogy be applied to ESP teaching combined with updated ICT (Information Communication Technology) to put forward a new format of education the "SPOC +Tradition Education".

21.2 Introduction of ESP

From the early 1960s ESP has grown to become one of the most prominent era of EFL teaching today (Thomas 2013). The end of the Word War II portended unparallel expansion in science, technology, and economy worldwide. English became the key to the international currencies of technology and commerce. Hutchinson and Waters (1987) gave three reasons for the emergence of ESP, the demands of a brave new world, a revolution in linguistics and a new focus on the learner. Today it is still a prominent part of EFL teaching (Anthony 1997). Johns and Dudley-Evans (2001) state that, 'the demand for English for specific purposes… continues to increase and expand throughout the world.' As Belcher (2006, p. 134) says ESP now encompasses an 'ever-diversifying and expanding range of purposes.' This continued expansion of ESP into new areas has arisen due to the ever-increasing 'glocalized' world (Robertson 1995). Flowerdew (1990) attributes its dynamism to market forces and theoretical renewal.

What exactly is ESP? Even today there is a large amount of on-going debate as to how to specify what exactly ESP constitutes (Belcher 2006; Dudley-Evans and St. John 1998; Anthony 1997). In the opinion of Hutchinson and Waters (1987), ESP is 'an approach to language teaching in which all decisions as to content and method are based on the learner's reason for learning'. To get a clearer picture, ESP can be likened to the leaves and branches on the tree of language and we can compare EGP to the roots of this tree of language (Thomas 2013). A significant aspect of language instruction at a tertiary level is learning English for a given purpose, with the specific aims of getting to know specialized vocabulary, increasing one's knowledge of the subject matter by reading in English and being able to use the language in the prospective profession or study areas by becoming prepared for some common situations such as carrying out higher level studies, going for an interview or conducting professional correspondence (Varnosfadrani

2009). ESP aims to provide learners with skills to read, write, listen, speak, and access information related to their professional study or specific job requirements. Following is a working definition for ESP:

> ESP is a learner-centered teaching method or idea, which is not a special language product (materials or methodology). ESP is based on the needs analysis of learners and stakeholders in a specific discipline, so as to determine to teach one or a few communication skills. The use of appropriate English of specific domain in the right occasions to communicate is focused in order to meet learners' professional study or specific job requirements. By Author

There is no single, ideal role description for an ESP practitioner. In addition to the routine tasks of a language teacher, the ESP practitioner may be required to deal with administrative, personnel, cross-cultural, interdisciplinary, curricular, and pedagogical issues that may be unfamiliar to ESOL teachers (Koh 1988; Waters 1994). This paper describes the characteristics of ESP as following: ESP learners are adults with intermediate or advanced English level, clear learning objectives. The content of teaching is based on the analysis of the learners needs in their specialized disciplines of work or study, combining specific disciplines to select or write appropriate materials, teaching single or multiple communication skills of the related disciplines. ESP is divided into EOP (Vocational English) and EAP (English for Academic Purposes), where EAP is still divided into EGAP (general academic English) and ESAP (professional academic English) (Jordan 1997).

21.3 A Brief Introduction of Knowles' Andragogy Theory

Malcolm Sherperd Knowles is a famous American adult educators. In 1967, he proposed the concept of "Andragogy". In 1996, Knowles was called "adult education evangelist" by U.S. adult education. Knowles' in-depth study and exploration on adult education led the interdisciplinary become an important part of the national education.

Knowles' Andragogy Theory is based on four assumptions which guide all adult education activities:

21.3.1 Adults' "Self-Directedness" Concept

Andragogy assumes that when a person grows up, his self-concept moves from one of total dependency to one of increasing self-directedness. An adult has a deep psychological need to be perceived as being self-directing. Any experience that they perceive as putting them in the position of being treated as children bound to trigger their resentment and resistance. (Knowles 1973) In order to realize adults' self-directedness, adult learners have to develop self-learning abilities, such as learn to make plans, observe, read, memory, and take notes in an effective way, etc.

21.3.2 Adults' Experience

This assumption is that as an individual matures he accumulates an expanding reservoir of experience that causes him to become an increasingly rich resource for learning, and at the same time provides him with a broadening base to which to relate new learnings. Andragogues convey their respect for people by making use of their experience as a resource for learning (ibid). Adults show diversified and personalized features on educational background, living environment, learning styles, hobbies, and interests, etc., so they have large heterogeneity in learning. Accordingly, the teaching methodology such as discussion, E-learning, simulation, team project, seminar, and other action-learning techniques can be adopt to better tap adult learners experience in learning.

21.3.3 Adults' "Readiness to Learn"

Andragogy assumes that adult learners are ready to learn those things they "need" to because of the developmental phases they are approaching in their roles as workers, spouses, parents, organizational members and leaders, leisure time users, and the like. Thus, to time learning experiences to coincide with the learners' developmental tasks is important (ibid). The changing tasks bring new learning opportunities, thus it is advisable to create or simulate roles through practically meaningful tasks for adult learners.

21.3.4 Adults' "Problem-Centered" Learning Orientation

This assumption is that adults want to apply tomorrow what they learn today, so their time perspective is one of immediacy of application. Therefore, they enter into education with a problem-centered orientation to learning. They would see as much more relevant a curriculum that is organized around the problem areas with which their work deals (ibid). Thus, it is advised that the curriculum be organized around problem-centered which can highly stimulate students' spirit in learning which has already been proved by Knowles' experiment in Boston University (ibid).

21.4 Application of Knowles' Andragogy Theory to English for Environmental Science

In Wenzhou University, ESP has been combined with different disciplines to meet students' future professional requirements in English. Thus, the curriculum concentrates on skill and field experience related English ability which requires problem-centered units rather than subject centered. Knowles' Andragogy Theory can well be applied in ESP.

What type of new education format can be adopted for the reform? Recently, higher education demonstrates new fundamental characteristics as accessibility and creativity. With the rapid progress in information technologies, MOOC (Massive Open Online Course) have opened a massive door in terms of online learning and making high-quality education free for all and open to the masses. The initial success of MOOCs is in the US, especially the impressive scale of student enrollment in the MOOC courses offered by elite schools such as Stanford and MIT. In China, top universities such as Peking University and Tsinghua University have put MOOC into action. Despite the fast growth of MOOC resources and infrastructure, accessibility of MOOC has not yet gained enough attention in China. This is because major MOOC providers, including universities and online platform vendors, are still busy with infrastructure development and MOOC course building, and have put accessibility aside (Wu et al. 2013). As Rodriquez declared, however powerful MOOC is, one of its major problems is that MOOC cannot correct the critical thinking work so that the students registered never have the opportunity to write papers or reports to have in-depth discussion with the professors. Another problem is that although face-to-face discussions between instructors and learners, integrated with online course contents, are developing in its initial stage, the effective combination of MOOC with the on-site course has no feasible and efficient way by far.

Taking advantage of the flexible educational process of MOOC, University of California-Berkeley Professor Armando Fox first coined the word APOC in 2013 to refer to a localized instance of a MOOC course that was in use in a business-to-business context. SPOCs support blended learning or flipped classroom learning, a current trend in education, which combines online resources and technology with the personal engagement between faculty and students that in-classroom teaching provides (Wikipedia). Harvard University announced incorporating SPOCs into its curriculum in the fall of 2013. Université catholique de Louvain has already proved a successful trier in implementing SPOCs with the course LFSAB1402 Informatics 2, a second-year bachelor university-level programming course taught to all engineering students (Combéfis et al. 2014).

Based on Knowles' four assumptions and the experience in Louvain, English for Environmental Science in Wenzhou University is under recasting into SPOC to promote the formation of students' self-learning ability and individualized learning strategies. The basic principles are as follows:

1. Adults' "self-directedness" Concept requires for self-learning abilities and a good command of self-learning process. Watson declares that people learn best that which they participate in selecting and planning themselves. In the SPOC platform, all the adult learners and lecturers are involved in the selection of practical themes for each unit and all the pre-questions and post-questions are posed after in-depth discussion. In this platform, students can learn on their own steps and practice learning strategies such as reading, discussion, taking notes etc.
2. Adults' experience emphasizes the precious value of adults' experiences, as Watson claimed how "ready" we are to learn is contingent upon adequate existing experience to permit the new to be learned. This will be fully realized in the reformed way of learning. First, the themes in SPOC are selected from real work or everyday life, which can activate students' previous passive learning experiences; second, there are abundant of learning materials in the platform for students to select and learn based on their special experience; third, students can choose tasks and the way to present them which are closer to their previous experience.
3. Adults' "readiness to learn orientation points out adults' learn better when their social responsibilities and obligations change. How "ready" we are to learn is contingent upon adequate significance and relevance for the learner to engage in learning activity. The best time to learn anything is when whatever is to be learned is immediately useful to us. In the "SPOC+Tradition Education" format, lecturers explain the purpose of tasks and use role plays, simulation etc. to embody real life or work situation to make in-the-spot experience which can inspire adult learners' motivation to learn.
4. Adults' "problem-centered" learning orientation declares that the problems/tasks assigned should be life-oriented and practical. Rogers (1969) points out that the facilitator helps to elicit and clarify the purposes of the individuals in the class as well as the more general purposes of the group. In the reformed format, theme related questions or practical problems are discussed and posed in and after class, students are asked to discuss in the discussion forum and present their tasks in various ways.

The recasted course was given in the second semester of the 2013–2014 academic years at School of Foreign Language in WZU. Two steps have been taken for the pilot study: a SPOC platform has been built with the uploading of the recast self-designed teaching material "English for Environmental Science;" Lecturers and students have been trained to use the SPOC and encouraged to try the new way of teaching and learning.

21.5 The Construction of College English SPOC

SPOCs can include video lectures, assessments (usually with instant feedback), interactive labs (usually with instant feedback), and discussion forums used in MOOCs (Wikipedia). Based on the basic models of MOOCs, the SPOC constructed in Wenzhou University is composed of six modules (see Fig. 21.1):

a. *Videos Lectures*: The videos are downloaded from the existing theme related resources such as the MOOC, youtube, etc. Multiple choices, numerical answers and short answers can be imbedded in the videos or after video shows to attract students' attention and better students' understanding of the videos. The benefits of Video lectures are proved by Deslauriers et al. (2011) who reported that active learning-based study improved learning by around 30 % higher in students' attendance, engagement and learning compared with traditional lecture-based learning (Fig. 21.2). Meanwhile, Notes-taking board will be inserted below the video window to facilitate students' learning.

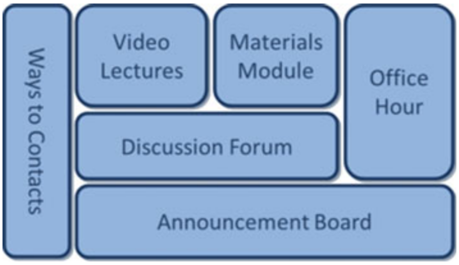


Fig. 21.1 Six modules of SPOC

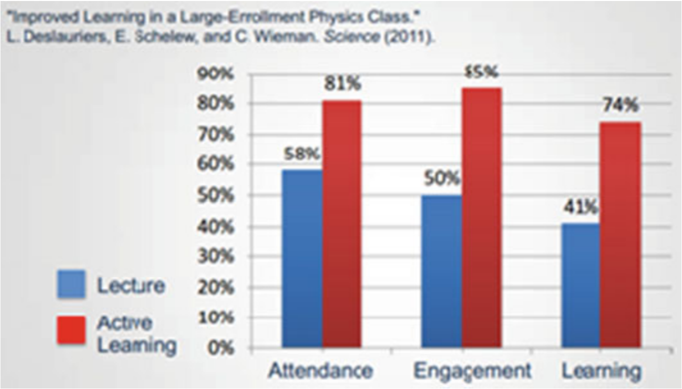


Fig. 21.2 Improved learning in a large-enrollment physics class (2011)

b. *Materials Module*: Lecturers as well as the students have the right and duty to contribute and complete the teaching materials which will be uploaded after expert's review. The materials should be real life-based, either be web links, electric books or articles, students' own works as ppt, reports, articles, videos, etc. "… learning is not merely through language, but with language". Teachers should endeavor to organize and make easily available the widest possible range of resources for learning (Rogers 1969).
c. *Discussion forum*: The forum is organized according to different theme related questions (pre-lesson questions, post-lesson questions). For example, if the lecturer gives five pre-questions to different students groups, the discussion forum will set five rooms, each student can choose different rooms in their favor to share opinions. Usually, a theme discussion lasts 2 weeks, after the deadline, the discussion forum will announce another new theme for discussion. As Houle (1972) declared, any design of education can best be understood as a complex of interacting elements, not as a sequence of events. In the discussion forum, students can raise any theme related questions, make statements, give suggestions and make comments.
d. *Office-hour*: In this virtual office-hour room, students can make an appointment once every 2 weeks with lecturers to have a video session or use Skype to exchange ideas, solve puzzles, or ask for help. Lecturers can also make time schedule to "meet" students to give instruction and check their progress.
e. *Announcement board*: Mainly each week, there will be messages published on the announcement board about theme related online reading lists or activities that will be organized after class. In this way, students can be monitored in making study plans.
f. *Contacts*: Ways to contact lecturers or teaching assistants such as e-mail, Skype, QQ, etc., are provided here for students to contact lecturers in a private way.

SPOC can provide learner flexibility and convenience, but how to provide a balance between flexible learning options and the high touch human interactive experience? It is known that many learners prefer the convenience offered by a distributed environment; meanwhile, they do not want to sacrifice the social interaction and human touch as in a face-to-face classroom. To better solve this problem, a new format of education "SPOC+Tradition Education" is applied. The updating pedagogy used in the new format of education is flipped classroom, with the belief that active learning approach can inspire the passion of learning. Students, through interaction with one another and share "ideas, hunches, queries… in the hope that these interactions will trigger other insights" since "knowledge is social in nature and constructed through a process of collaboration, interaction and communication among learners in social settings" (DeWaard et al. 2011). In this type of classroom, students play an important role in presenting their learning outcomes such as seminars, presentation, speeches, role play, and short video shows which are recorded before class. Lecturers play roles as facilitator, organizer, inspiration, learning partner, and coordinator.

21.6 Structure and Timeline of the Course

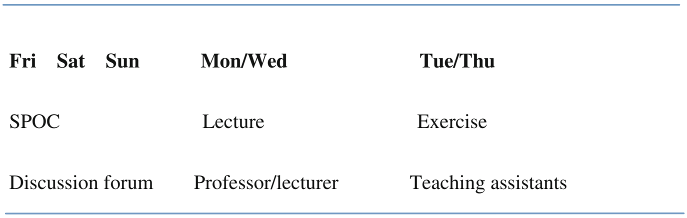
The developed SPOC is based on English for Environmental Science, which is realized by transdisciplinary cooperation between English lectures group and professors from environmental protection department. The course is taught to 119 second-year bachelors in Environmental Science in 2013. To integrate the SPOC with the newly set course, the first step is to split the course material in two tracks: one is realized by SPOC, the other is taught with a traditional course. The SPOC track contains the basic information of the course and the updated materials downloaded from theme related MOOCs or Youtube like on line resources. Meanwhile, the traditional course brings advanced concepts and gives students opportunities to present their learning outcomes.

Houle (1972) once identifies the task of the educator to manage which include design a suitable format:

a. Learning resources are selected.
b. A leader or group of leaders is chosen.
c. Methods are selected and used.
d. A time schedule is made.
e. A sequence of events is devised.
f. Social reinforcement of learning is provided.
g. The nature of each individual learner is taken into account.
h. Roles and relationships are made clear.
i. Criteria for evaluating progress are identified.
j. The design is made clear to all concerned.

Based on Houle's format, the course is split into cycles that follow the same pattern (see Fig. 21.3). A cycle presents a week's schedule, and the course lasts 16 weeks. What's more, each cycle corresponds one SPOC lesson.

The first part of the cycle is the SPOC. First, the students have to study the materials in SPOC and finish the selected pre-question raised by lecturers before class. This PBL aims to provide acquisition of information based on facts. Thus, all the problems are chosen out of the real world. Students can present their problem solving results in any suitable style such as PPT, video show, etc. The SPOC track is implemented with the Instructure Canvas; it consists of videos presenting the



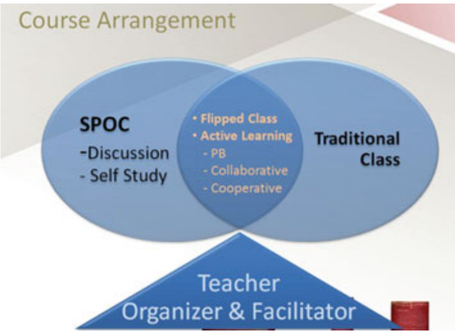
Fri Sat Sun	Mon/Wed	Tue/Thu
SPOC	Lecture	Exercise
Discussion forum	Professor/lecturer	Teaching assistants

Fig. 21.3 Organization of a week for the remastered course

basic concepts, updated information and exercises of various kinds which can be downloaded from the existing web resources such as the MOOCs and youtube or the lecturers can create their own videos. All the videos are modulated into short units such as 5–10 min. Exercises are interleaved with videos, which can be proposed in or after each video or groups of videos. Exercises can be classical multiple choices or short questions.

After finishing working on the SPOC track, students have to attend an hour and a half lecture. The lecture is split in two parts. The first part is students' presentation time where students present their cooperation results of the pre-question selected before class. Each presentation has the time limit of around 5 min which can be realized in different types such as video show, PPT, debate, role play, or speech depending on the content of their work. Each group's presentation is followed by another group's comments according to the preset evaluation criterion. The teacher plays a role as organizer and facilitator to push forward the presentation part. Teaching mentors fulfill this role by monitoring discussions, asking questions, helping the resolution of occasional conflicts, enabling the participation of each group member to classroom discussions, giving examples when required, preventing scatter of discussions and making evaluations (Duffy and Cunningham 1996; Rhem 1998; Greenwald 2000; Posner and Rudnitsky 2001). In this way, teachers play an important role in helping to make globally distributed materials culturally relevant and meaningful. The second part is dedicated to present advanced or key concepts and explain the difficult points on the SPOC track.

Finally, the last activity of the 1 week cycle is an hour and a half exercise session. In the first class, students are supervised by lecturer or teaching assistants to make sure they have understood the theoretical aspects on the SPOC and some common mistakes of the SPOC exercises are explained. In the second class, students receive supplemental on-paper or oral exercises covering the basic language skills such as translation, listening, and reading. Or, they will be asked to write an in-class report or short article on the theme after a week's study including SPOC and class presentation. After finishing the 1 week cycle, post-class questions are raised for students' further study on the theme (see Fig. 21.4).



Course Arrangement

SPOC
-Discussion
- Self Study

• Flipped Class
• Active Learning
- PB
- Collaborative
- Cooperative

Traditional Class

Teacher
Organizer & Facilitator

Fig. 21.4 The Andragogy of "SPOC+Tradition Education" Format

It is important to change students' and lecturers' traditional concepts on teaching and learning and adapt to an active leaning approach following the flipped classroom pedagogical model. Thus, in the first week of the semester, introduction lectures are given to the all the students and teachers involved. The purpose is to introduce the organization, function, and evaluation of the new pedagogy model, and to explain the structure and syllabus of the new format.

The evaluation of the remastered course contains two parts. The first part is SPOC evaluation which counts a small part because it's difficult to verify students' performance on line. Besides, since it's a brand new learning concept and experience for students and teachers as well, a higher proportion in the final score may cause nervous and inner resistance. The second part is the evaluation of traditional course which includes in-class group presentation and papers or reports that followed a midterm exam and a final exam.

21.7 Technological and Social Integration

The successful implementation of the remastered course needs the co-development in technical aspect and the social aspect as well (Combéfis et al. 2014). First, since the technics in E-learning is quite mature, experts can easily succeed in providing exercises that can be assessed automatically and give student feedback instantly. We use the Instructure Canvas to satisfy both requirements. InstructureCanvas is an online learning management platform that can automatically grade programs while providing intelligent and relevant feedback. SecondS, social aspect is an essential part of SPOC development. For one thing, the new concept of active learning by SPOC should be spread and rooted among lecturers, administrative staff and students. For another, to build the SPOC, a group of lecturers and technical experts are organized to in charge of selecting or creating videos and the corresponding exercise for SPOC. The tools needed are a webcam, a good-quality microphone and Camstudio software. Thirdly, two teaching assistants are needed to animate the discussion forum, to oversee and clarify the discussion

Meanwhile, to approve students' engagement in using the recorded material properly to ensure a correct and responsible active learning process, Pinvox—"Personal Identification Number by Voice" (Vox in Latin) is applied. It is a prototype algorithm-m that aims to guarantee that scholars have followed, i.e., listened to and watched a complete recorded lecture with the option of earning a certificate or diploma of completion after attending courses virtually. It aims to providing a way to be able to tell whether someone watched a video completely or listened to a whole audio file. Canessa and Logofatu (2013) reported that the student's engagement becomes higher since he/she gets more and more involved with the self-study as the embedded Pinvox sequences capture their attention. These benefits are coupled to those in the flexibility of study and in the increase of student's responsibility.

Since the new active learning in SPOC needs cooperation from all aspects of school or university in this case, the people involved should collaborate closely and frequently to be in charge of the feedback about the current situation and coordinate all elements involved.

21.8 Evaluation of the Remastered English for Environmental Science

The construction of the remastered course is still under the way of perfection after its prototype being finished by the end of fall 2013. Thus, only one batch of students majored in Environmental Science has been taken part in the "SPOC+Tradition Education" format learning. The data-based evaluation of the remastered course will not be complete until the first batch of students completely adapt to the SPOC model which will take at least 1 year. However, we have made a questionnaire survey to 119 subjects majored in Environmental Science in Wenzhou University to learn about their attitudes toward the SPOC model and its impact on them.

The "SPOC+Tradition Education" format combines different forms of active learning including collaborative learning, cooperative learning, and problem-based learning (PBL) through engaging students into all types of activities such as group discussions, problem solving, role plays, and simulation games. The benefits to using such activities are many. As the Center for Teaching and Learning in the University of Minnesota pointed out that the benefits include improved critical thinking skills, increased retention and transfer of new information, increased motivation, and improved interpersonal skills. Prince (2004) also declared that active learning enhances academic achievement, produces positive student attitudes, improves the long-term retention, better study habits, and develop enhanced critical thinking and problem-solving skills. To find out the active learning result of The "SPOC+Tradition Education" format, a questionnaire was made based on the above-mentioned elements of active learning which include learning retention, learning motivation, critical thinking, problem-solving skills, better learning habits, Interpersonal skills, and Creative thinking (see Table 21.1).

A very small portion (less than 3 %) reported that they are dissatisfied with the result of active learning in the new format of education. This finding is consistent with Bonwell and Eison (1991) who conclude that active learning leads to better student attitudes and improvements in students' thinking and writing. Felder et al. (2000) include active learning on their recommendations for teaching methods that work. Akınoğlu and ÖzkardeşTandoğan (2007) pointed out that by means of PBL (one form of active learning adopted in the new format of education), some attitudes of students in relation to such areas as problem solving, thinking, group works, communication, information acquisition and information sharing with others are affected positively.

Table 21.1 Active-learning satisfaction questionnaire ($N = 119$)

Questions	Completely satisfied (%)	Mostly satisfied (%)	Neither satisfied nor dissatisfied (%)	Mostly dissatisfied (%)	Completely dissatisfied
Leaning retention	21	49	30	0	0
Learning motivation	11	50	38	1	0
Critical thinking	13	39	47	1	0
Problem-solving skills	6	44	48	2	0
Better learning habits	9	43	45	3	0
Interpersonal skills	12	43	44	1	0
Creative thinking	16	52	31	1	0

Note Measured on a five-point scale with response categories ranging from (1) completely dissatisfied to (5) completely satisfied. (** $p < 0.01$.)

The highest proportion (around 50 %) reported that the new format of education helps enhance their learning retention, motivation, and creative thinking. This result is consistent with Fredericksen and Berry who suggest that collaboration (one form of active learning) is particularly effective for improving retention of traditionally underrepresented groups. Gehringer and Miller (2009) declare that students who engage in active learning activities are more attentive during class.

In general, students are satisfied with the new SPOC model in improving their active learning ability. For the open question part, students listed their difficulties in "SPOC+Tradition Education" learning, among which cooperation in tasks comes the first. The possible explanation is that first, the traditional Chinese way of learning encourages solo study since primary school, thus, it takes time for students to learn to cooperate in collaborative learning; secondly, students live in the digital world lack interpersonal skills, thus, it takes efforts to learn to make compromise and distribute individual tasks among group members. Lecturers have put more effort into helping and guiding students to build the sense of co-study and to process interpersonal skills. Besides, students also mentioned that they have difficulty in processing and digesting the huge information input during the class presentation and the massive material in the SPOC. On the one hand, this reflects the innovated mode of delivering curriculum has a challenging higher requirement on students' learning strategies and leaning habits as well; on the other hand, the new format of education has a higher requirement on teachers' classroom management ability and curriculum design foresight.

21.9 Conclusion

The new "SPOC+Tradition Education" learning format based on Knowles' Andragogy Theory combines different forms of active learning including collaborative learning, cooperative learning, and problem-based learning (PBL) has shown its advantages in the pilot study. From the survey, we can conclude that the new format of education helps enhance students' learning retention, motivation, and creative thinking in a positive way. It is observed that students taking the remastered course show the passion that has been buried for ages in traditional way of teaching; besides, they reported that they have improved problem-solving skills, transferable workplace experience, better communication skills, improved leadership ability, and critical thinking ability as well. Although a great number of the students reveal their weakness and worries in interpersonal relationship, cooperative learning, and proper learning strategies, and the teachers complain about the higher requirements on class management ability and enormous time spent in curriculum design, it is an essential stage of education reform to prepare students' adaptability in fast-changing environments.

The perfection of the SPOC from an existing course is time and energy costly; however, the experiment can be taken as a trying step in taking advantages of streaming technologies to create a mixed mode of delivering curriculum for the benefits of the students, teachers and universities.

References

Akınoğlu, O., & ÖzkardeşTandoğan, R. (2007). The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. *Eurasia Journal of Mathematics, Science & Technology Education, 3*(1), 71–81.

Anthony, L. (1997). ESP: What does it mean? Why is it different? On Cue. Retrieved November 15, 2008 from http://www.antlab.sci.waseda.ac.jp/abstracts/ESParticle.html.

Belcher, D. (2006). English for specific purposes: Teaching to perceived needs and imagined futures in worlds of work, study and everyday life. *TESOL Quarterly, 40*(1), 133–156.

Bonwell, C. C., & Eison, J. A. (1991). Active learning: Creating excitement in the classroom. ASHEERIC Higher Education Report No. 1. Washington, DC: George Washington University.

Canessa, E., & Logofatu, B. (2013). Pinvox method to enhance self-study in blended learning: Experiences at University of Bucharest. *iJET 8(*2):53–56.

Combéfis, S., Bibal, A., & Roy, P. V. (2014). *Recasting a Traditional Course into a MOOC by Means of a SPOC*. European MOOCs Stakeholders Summit 2014 (pp. 205–208). Switzerland: Lausanne.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 332*(6031), 862–864.

DeWaard, I., Abaljian, S., Gallagher, M. S., Hogue, R., Keskin, N., Apostolos, K., & Rodriquez, O. C. (2011). Using MLearning and MOOCs to understand chaos, emergence and complexity in education. *The International Review of Research in Open and Distance Learning, 12*(7), 95–112. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1046/2043.

Dudley-Evans, T., & St. John, M. J. (1998). Developments in English for specific purposes: A multi-disciplinary approach. Cambridge: Cambridge University Press.

Duffy, T. M. & Cunningham, D. J. (1996). Constructivisim: implications for the design and delivery of instruction. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology*. New York: Simon & Schuster Macmillan.

Felder, R., D., Woods, J. S., & Rugarcia, A. (2000). The future of engineering education: II. Teaching methods that work. *Chemical Engineering Education, 34* (1), 26–39.

Flowerdew, J. (1990). English for specific purposes: A selective review of the literature. *ELT Journal, 44*(4), 326–337.

Gehringer, E. F., & Miller, C. S. (2009). Student-generated active-learning exercises. *Proceedings of the 40th ACM technical symposium on Computer science Education Chattanooga, TN* (pp. 81–85). USA: ACM.

Greenwald, N. L. (2000). Learning from problems. *The Science Teacher, 67*(4).

Houle, Cyril O. (1972). *The design of education*. San Francisco: Jossey Bass.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centered approach*. Cambridge: Cambridge University Press.

Johns, A. & Dudley-Evans, T. (2001). English for specific purposes: International in scope, specific in purpose. *TESOL Quarterly*, *25*(2), 297–314.

Jordan, R. R. (1997). *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Knowles, M. (1973). *The adult learner: A neglected species*. Houston: Gulf Publishing Company.

Koh, M. Y. (1988). The changing role of the ESL teacher and its implications for ESP. In M. Tickoo (Ed.), *ESP: State of the art. Anthology Series* (Vol. 21, pp. 74–79).

Posner, G. J., & Rudnitsky, A. N. (2001). *Course design*. New York: Addison Wesley Longman, Inc.

Prince, M. (2004). Does active learning work? A review of the research. *Engineering Education, 93*(3), 223–231.

Rhem, J. (1998). *Problem-based learning: An introduction*. The National Teaching & Learning Forum, 8, 1. USA: Oryx Press.

Robertson, R. (1995). Glocalization: Time-space and homogeneity-heterogeneity. In M. Featherstone, S. Lash & R. Robertson (Eds.), *Global modernities* (pp. 24–44). London, Sage.

Rogers, C. R. (1969). *Freedom to learn*. Columbus: Ohio. Merrill.

Strevens, P. (1977). *New orientations in the teaching of English*. Oxford: Oxford University Press.

Tan, O. S. (2013). *Thinking skills, creativity and problem-based learning*. Singapore: Temasek Polytechnic. Retrieved from http://www.tp.edu.sg/files/centres/pbl/pbl_tanoonseng.pdf.

Thomas, M. (2013). The significance and relevance of English for specific purpose [ESP]. *Confluence*, 159–162.

Varnosfadrani, A. D. (2009). Teaching English for specific purposes. In R. Reinelt (Ed.), *Into the next decade with (2nd) FL teaching* (pp. 181–201). Japan: Rudolf Reinelt Research Laboratory EU Matsuyama.

Waters, A. (1994). ESP things fall apart. In R. Khoo (Ed.), *LSP. Problems and prospects* (pp. 1–14). Singapore: SEAMEO Regional English Language Centre.

Wu, W., Hu, C., & Zhang, K. (2013). MOOC and accessibility in China. Accessible E-Learning Online Symposium of 16 December 2013. http://www1.umn.edu/ohr/teachlearn/tutorials/active/what/.

Appendix

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014AU003

An Evaluation of the Tampa Scale for Kinesiophobia as an Estimate of Fear-Related Movement for People with Dystonia

Presenter
Dr Ian Blackman
A Lecturer within the School of Nursing & Midwifery at Flinders University
Adelaide, South Australia

Dr Lynley Bradnam,
A Senior Lecturer in Physiotherapy at Flinders University in Adelaide
Australia

Mr Lynton Graetz
A Student in the Department of Psychology at Flinders University
Adelaide, Australia

Background

Dystonia is a neurological movement disorder characterized by sustained or intermittent muscle contractions causing abnormal, often repetitive, movements, postures or both. Dystonia impacts on psychosocial function with one psychological aspect that is poorly understood: fear of movement. The Tampa Scale of Kinesiophobia (TSK), originally used with people with low back pain, has been adapted for other musculoskeletal disorders but has not been used to assess (fear-related) movement behaviors associated with people with dystonia. Past studies that have adapted the TSK tool have used a range of analytical methods to determine its psychometric qualities but with mixed results. In this current study, Rasch analysis was used to determine the validity and reliability of the TSK for Dystonia survey.

Aims

The primary objective of this research is to assess the psychometric properties of the questionnaire using Rasch analysis and to adapt the Tampa Scale of Kinesiophobia (TSK) for use with people with Dystonia (TSK-Dystonia).

Keywords Dystonia • Kinesiophobia • Rasch analysis

Sample

Participants: 156 attempted the survey with 132 completing the whole survey (84.6 %). Of the 132 completions 108 (82 %) were female respondents and 24 (18 %), were male. All were diagnosed with Dystonia by a medical officer, were English speaking and lived in Australia, New Zealand, USA, or in Europe. Participants' mean age was 48 (sd = 12 years).

Method

The original 17 item TSK survey was adapted for Dystonia using focus groups of patients and health professionals with experience treating dystonia. The online questionnaire was distributed online via networks of Dystonia groups. Survey results were analyzed using the partial credit version of the Rasch model using Conquest and Winsteps software, to specifically explore questionnaire item fit, scale threshold organization, and differentiated item functioning.

Results

Rasch analysis identified four survey items as not demonstrating unidimensionality, with another two items having poor item threshold capacities. These items were deleted from further analysis. Participant gender and type of diagnosed dystonia as two additional variables produced unwanted variance (DIF), necessitating the possible removal of an additional two other survey items. The remaining TSK survey items were able to demonstrate a reliable conjoint hierarchical scale, based on differing consensus estimates of participants and their differing beliefs about dystonia.

Conclusions

Rasch analysis found that a shortened form of TSK-dystonia survey could be suitable to identify factors of fear-related movement in people with dystonia however, further survey adaptation is recommended so participants’ consensus at the lower end of the scale (ability for disagreement) can be more fully estimated.

PROMS2014AU004

The Effectiveness of the MISSCARE Tool in Determining the Frequency, Type, and Reasons for Missed Nursing Care: A Psychometric Evaluation

Presenter

Ian Blackman
EdD. BEd. MEd. RN. RPN. GradDip (HealthCouns),
School of Nursing and Midwifery, Flinders University
Adelaide, South Australia, Australia

Eileen Willis
Ph.D., BEd, MEd (Sociology), School of Health Sciences, Flinders University
Adelaide, South Australia, Australia

Luisa Toffoli
Ph.D., RN, MN Lecturer, School of Nursing, University of South Australia
Adelaide, South Australia, Australia

Julie Henderson
Ph.D., BA (Hons), Grad Dip (Information Studies),
School of Nursing and Midwifery, Flinders University
Adelaide, South Australia, Australia

Patti Hamilton
Ph.D., MS (Nursing) BS (Nursing), Texas Woman's University
Denton, Texas, United States of America

Claire Verrall
RN, BN, MN, School of Nursing and Midwifery, Flinders University
Adelaide, South Australia, Australia

Elizabeth Abery
BHSc (Honours), School of Medicine, Faculty of Health Sciences, Flinders University
Adelaide, South Australia, Australia

Clare Harvey
Ph.D., RN, MA, Eastern Institute of Technology
Napier, Hawke's Bay, New Zealand

Aims

The MISSCARE tool has been historically used to estimate what types, frequency, and rationales behind episodes of missed nursing care by using two Likert scales.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Rasch modeling has not previously been employed to explore the psychometric properties of the two MISSCARE scales. The major aim of the study is to use the Partial Credit Model to ascertain if the MISSCARE tool meets Rasch modeling prerequisites.

Sample
Three hundred and eighty nine qualified South Australian nurses employed within the private and public health care sectors completed the MISSCARE tool, online.

Method
A nonexperimental comparative approach was adopted and the two scales used to estimate frequency of omitted nursing care (24 items containing three thresholds) and why nursing care was missed (17 items with four thresholds) were analysed using the Partial Credit Model contained within the Conquest and Winsteps programs.

Results
One survey item contained within each of the scales estimating frequency of missed nursing care and why nursing care was missed, failed to meet the unidimensionality requirements of the Rasch model. Six items were seen to produce unwanted variance (DIF) related to participants' site of employment and gender. The two revised surveys produced acceptable item and person fit statistics when re-examined after poorly fitting items were removed.

Conclusions & Significance of findings
The survey items and person estimates of both scales used once revised, are fitting the measurement requirements for Rasch modeling. For the type and frequency of missed nursing care survey, the Item Separation Index (ISI) was 3.91 (reliability 0.94) and the Person Separation Index (PSI) was 3.77 (reliability 0.94) respectively, indicating a very good reliability index for the survey items. Similarly, the attribution scale's reliability indicating why nursing care was missed demonstrated very acceptable reliability also, with an ISI and PSI of 2.54 (reliability 0.87) and 6.71 (reliability 0.98). Differentiated item functioning (survey item bias) in the revised surveys indicated that survey items will not produce unwanted items bias in relation to participant gender and work place settings. Both scales therefore demonstrate a very acceptable capacity for data replicability, should the data be retested and the items used to measure the frequency of and the reasons for reported missed nursing care confirm their dependability for measurement. Results of the person-item maps of both scales provide researchers and managers with hierarchical linear structures readily identifying the complexities of nursing care (based on nursing care omission) as well as measures of attribution for why nursing care is missed which will influence decision making and the planning of contemporary nursing practices and the allocation of its associated nursing care resources.

PROMS2014CN003

Recasting College English Course into a SPOC—A Pilot Study in Wenzhou University

Presenter
Xiaoshu Xu, Ph.D Candidate
Wenzhou University
E-mail: xuxs1016@126.com

Xiuyu Gu, Ph.D Candidate
Macau City University
E-mail: 107078151@qq.com

Higher education is in the early stage of a revolutionary transition which calls for the creation of a new Adaptive Learning System to promote active learning by students. In china, College English course is undergoing a significant reform toward English for General Education, ESP (English for Special Purpose) and EAP (English for Academic Purpose). While, in Wenzhou University, the reform tendency is to built a groups of English outward bound courses. To facilitate the reform, the author, together with a group of colleagues, start to recast one of its new course Environmental English into a "SPOC+Tradition Education" format to create a new type of learning experience for the students exposed in digital, interconnected world. The approach has two steps. The first step is to construct a SPOC platform and remaster the course with the help of transdisciplinary professors and information technology experts. The second step is to spread the new concept of SPOC for the on-site course learning among the teachers, administrators and students as well as to undertake practical training for them. Altogether, 119 students of first-year bachelor majored in Environmental Science in Wenzhou University have taken part in this pilot study. The result of the questionnaire survey shows that the majority of the subjects is satisfied with the remastered course teaching, and believes that the new format of learning can inspire their learning interests, improve their learning strategy, and help them adapt to collaborative learning model. The data-based evaluation of the study efficiency will not be complete until the SPOC is completed. The purpose of conducting such a study is to explore the feasibility of using SPOC to transform traditional lecture-based teaching. The research report was submitted to the concerned department of higher education for reference.

Keywords College English • SPOC • Pilot study • Questionnaire survey

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

POROMS2014CN008

Involving Test Stakeholders in Test Validation: An Investigation into Test Candidates' Attitudes to the Versant English Test

Presenter
Jinsong Fan (Ph.D.)
A lecturer within College of Foreign Languages and Literature,
Fudan University
Shanghai, China
E-mail: jinsongfan@fudan.edu.cn

Background

Current validity frameworks in language testing and assessment require that multiple strands of validity evidence should be collected, weighed, and combined to form into a coherent and convincing validity argument to support test score interpretation and use (see Bachman 2005; Bachman and Palmer 2010; Kane 1992, 2001). Since test candidates are the most important stakeholders in any assessment situation, it is essential to investigate their attitudes to and views of a test as an important component of test validation.

Aims

This study investigated test candidates' attitudes to the Versant English Test (VET), a spoken English test developed by Pearson, through addressing the following three research questions:

(1) What is the pattern of test candidates' attitudes to the VET, and what are the sources of the positivity and negativity in test candidates' reported attitudes?
(2) Is there any difference in test candidates' reported attitudes that are related to their gender and academic background?
(3) What is the relationship between test candidates' reported attitudes and their test performance?

Method and Sample

To adequately address these three research questions, a mixed-method design was adopted. The participants in this study were 125 Chinese candidates who had just taken the VET in March, 2014. Research data were collected through a structured questionnaire consisting of 36 items on six-point Likert scale of agreement. In addition, 25 candidates participated in the follow-up interviews. A number of statistical analyses were adopted to process the questionnaire data including exploratory factor analysis (EFA), reliability estimates, multivariate analysis of variance (MANOVA), and stepwise regression analyses. EFA extracted five attitudinal factors, explaining 60.91 % of the common variance.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Results

Descriptive statistics at the factor and item levels demonstrated that test candidates' attitudes to the VET were on the whole positive, and they generally believed that the VET was a good indicator of their spoken English ability. MANOVA results indicated that gender and academic background did not affect test candidates' reported attitudes to the VET. Stepwise regression analyses showed that test candidates' perceived difficulty of the open-ended questions in the VET was the only factor which successfully entered the regression model, explaining about 5.3 % of test score variance.

Conclusions

The results of this study have lent important evidence to the validity of the VET. In addition, this study has also further confirmed the necessity and importance of involving test stakeholders in language test validation.

PROMS2014CN009

Investigation and Analysis of Wenzhou New Generation Returnees' View on Marriage and Love

Presenter
Xiuyu Gu, Ph.D Candidate
Macau City University
E-mail: 107078151@qq.com

Xiaoshu Xu, Ph.D Candidate
Wenzhou University
xuxs1016@126.com

A radical change is underway in the way Chinese look at dating and marriage in the information age, especially for the Wenzhou New Generation Returnees, who present a generation of blended culture and pioneer in openness. Based on a survey of 500 Wenzhou new generation returnees and local youth, married and single, this paper analyzes the modern situation, trends, and opinions of marriage and love of the former through comparison. The research finds out that the new generation returnees focus on pursuing advanced studies, and their careers are close to the mainstream society. However, in choosing a spouse, they manifest a liberal concept while behave conservatively. They support flash marriage and premarital sex while mainly hold negative attitude toward having affairs; they strongly favor "Wenzhou natives", a significant tendency towards "localized spouse selection"; they are keen on blind dates, accepting premarital birth horoscope check, and their parents have significant intervention rights on their marriage. In addition, a vast majority of them support the "cooperative marriage", and they strongly believe consistent personal values, characters and habits overweigh love in choosing spouse.

Keywords Wenzhou new generation returnees • View on marriage and love • Questionnaire survey

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN010

Developing a Teacher-Rated Early Childhood Developmental Scale

Presenter
Xiaoting Huang, Kexin Guan
China Institute for Educational Finance Research, Peking University
Beijing, China

Background

Early childhood education has gained increasing attention in China in recent years. However, scientific tools measuring children's development are scarce, and most of them were difficult to implement.

Aims

This study attempted to design a teacher-rate dearly childhood developmental scale to better inform teachers and parents the children's status. Altogether, there were 37 multiple-choice items measuring five dimensions, including health and fitness, language, social psychology, math and science, and arts.

Keywords Early child development scale • Reliability • Validity

Sample and Method

Teachers chose the category that best describes the skills and knowledge a child demonstrated based on their daily observations. 118 children from 32 classes in 4 kindergartens participated in the study. The data were calibrated using a multidimensional RCML model.

Results

Our results showed that the five dimensions are relatively stable with reliabilities ranging from 0.66 to 0.90. The five dimensions were also highly correlated. In addition, the instrument demonstrated cross-age validity as the ratings on all five dimensions improved with the increase of age. Moreover, we collected data from parent and teacher surveys to investigate the external validity of the instrument.

Conclusions

The analysis showed that parents' educational background and teacher's creativity had relatively large positive effects on these children's cognitive and psychological developments. These results indicated that the reliability and validity of this instrument were satisfactory. Kindergarten teachers could use this scale easily to assess the children and adjust their teaching agenda accordingly. The major drawbacks of the instrument were that items in math and science dimension were

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

too difficult than we desired and the range of item difficulty was not large enough. Future research can focus on developing more items in this dimension. Finally, the more samples of typical behavior for each scoring category will be helpful for teachers to rate the children more precisely.

PROMS2014CN015

Developing and Validating a Tool to Measure Foreign Language Aptitude of Chinese FL Learners: A Pilot Study

Presenter
Lanrong Li, Ph.D in Applied Linguistics
National Institute of Education Sciences
Beijing, China
E-mail: withorchid2006@sina.com

Background

Foreign language (FL) aptitude has been accepted as one of the most important individual variables in second-language acquisition. FL aptitude tests have been used for a variety of practical purposes like prediction, selection and diagnosing language learning difficulties. Despite the significance of aptitude and the potential values of aptitude tests, there has no been valid and reliable FL aptitude test for Chinese FL learners.

Aims

This study aims to develop a pilot and validate tool to measure FL aptitude of Chinese FL learners aged 16 and above by applying the Rasch model.

Keywords Foreign language aptitude • Rasch model • Validity • Test development

Sample

158 students from two senior high schools and one university in Beijing participated in the study.

Method

Test development was guided by Carroll's (1965, 1981) aptitude theories and the test was modeled after well-known aptitude tests abroad. Six subtests were developed to measure Carroll's (1965, 1981) four aptitude components, which are phonetic coding ability, grammatical sensitivity, inductive language learning ability and rote memory respectively. In order to make the test more suitable for classroom administration in the two senior high schools, two versions of the test were designed: Test A was for college students while Test B was for senior high school students. The aptitude test was administered to the students in fall 2012. The Rasch dichotomous model (Rasch 1960; Wright and Stone 1979) was used to examine the internal construct validity of the test. The external validity of the test was examined by studying the relationship between the senior high students' Rasch scores on the aptitude test and their English achievement test scores.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Results

Rasch analyses showed that on the whole the test was highly reliable and of appropriate difficulty for the students, and that each subtest showed good fit to the model and all the subtests except for the subtest of Number Learning showed unidimension. However, a small number of misfitting items were identified in some subtests, the subtest of Paired Words was found to be too easy for the students, and one subtest measuring phonetic coding ability was not correlated with other subtests or the senior high school students' English achievement.

Conclusion

Overall, the test was of high reliability and validity. However, the subtests measuring rote memory were found to have a narrow range of item difficulty and one subtest measuring phonetic coding ability may not be a valid measure of this component. These subtests were correlated with each other, supporting the structural validity of the test. The results also suggest that the test could predict the English achievement of senior high school students, though different aptitude components seem to play different roles in language learning different stages.

PROMS2014CN016

Quality Analysis Based on Rasch Model of PIRLS2006 Sample Test for Guangxi Subjects

Presenter
Jing Yu
Postgraduate of Guangxi University

Jing Gong
Postgraduate of Guangxi University

Dehong Luo
Professor of Guangxi University
E-mail: headfar@126.com

Background

Chinese mainland fundamental education attaches great importance to reading assessment, whose model is empirical and far away from scientific assessment theory and tool.

Aims

In order to estimate the current situation of Guangxi students' reading literacy and improve application of scientific assessment tool, this research, in 2012, stratified 10-11-year-old 860 pupil samples from 37 classes, 12 elementary schools in 9 cities of Guangxi to take the sample test of PIRLS2006 (Program of International Reading Literacy Studies), and Bond & Fox steps 1.0.0 software and Rasch Model are used to analyze it.

Keywords PIRLS2006 sample items • Rasch model • China's Guangxi • Test quality

Methods

The hypotheses are as follows: (1) PIRLS2006 sample test is a high quality test tool; (2) reading literacy level of Guangxi samples are lower than the difficulty of test; (3) Guangxi samples' reading literacy performances are poorer in integration and interpretation, criticism and evaluation.

The following steps are taken to test the hypotheses. Multidimensionality Investigation is firstly taken to determine the Rasch Model application suitability. It's found that Correlation Coefficient of a large number of items is [−0.4, 0.4]. The other examining steps are Wright map, overall statistics, model-data fit, error statistics, bubble diagram, Differential Item Functioning and item characteristic curve.

Results

The followings are research results: (1) distributions of both reading level and items are abnormal, with great differences in reading literacy performance and items

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

difficulty. Higher-than-average-level pupils are a little more than the lower ones, mostly distributing in one standard deviation. Items difficulty of above-average and under-average covers each type of item, with balanced distribution; (2) Correlation Coefficients of all items are positive, the test reliability is 0.99, item difficulty error is 0.08. Although the outcome of model fit of most of the items is good, with low error of estimation, a few items don't fit well; (3) Guangxi pupils' reading literacy performance is −3.26 logit, much lower than the difficulty of PIRLS2006. The reading literacy error is 0.47. The items that do not fit well with the model are mostly those concerning the literacy of integration, interpretation, criticism and evaluation; (4) No serious DIF between genders exist.

Conclusions

Conclusions of this research demonstrate that some particular items have not achieved the expected assessment goal, and should be adjusted. The reading literacy performance of Guangxi pupils, especially in the literacy of integration and interpretation, is lower than most of the PIRLS2006 assessed countries, ranked the 34th. We suggest that, in order to set up scientific and rigorous reading literacy assessment system, as a pilot study, the international reading literacy assessment tool, theory and framework be applied in Chinese mainland fundamental schools and Rasch Model be used as a tool to examine the quality of the test.

PROMS2014CN017

Backwash Effect of English Testing on English Teaching for Adults

Presenter
Yuhuan Gong
City East District Vocational College
Beijing, China
E-mail:gongyuh2004@163.com

At present, there have been a great many researches on backwash effect of English testing on English teaching in Universities, few of which involves English teaching in Adult colleges. This paper discusses the quality criteria of the English testing like the validity, the reliability and so forth, referencing to the research and the theory on the relationship between language teaching and testing at home and abroad. The paper also reveals the backwash effect of English testing on English teaching in the adult college. Meanwhile, reform proposals to English testing have been made in order to bring the positive function of English testing into play and to improve English language testing scientifically.

Keywords Test • Backwash effect • English teaching for adults

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN018

Enlightenment of TOEFL IBT on English Teaching and Its Evaluation System Reform in the Adult Colleges

Presenter
Haihong Li
City East District Vocational College
Beijing, China
E-mail: lhhjoy@126.com

Abandoning the structuralism testing patterns, TOEFL IBT with the Internet as the medium, aims to thoroughly test the candidates' comprehensive English language ability of listening, speaking, reading, and writing and the ability of using the language. In this sense, TOEFL IBT redefines the way English language is learned and measured and perfectly embodies the communicative language competence testing concept and communicative language teaching idea, both of which have a significant implication on English teaching and its evaluation system in the adult colleges. This paper discusses the current situation of English testing system and its existing problems in the adult colleges. Meanwhile, some reform proposal have been made to strengthen the English teaching reform and fully realize the teaching objective in the adult colleges which gives priority to the practicality and targets at fostering students' comprehensive English communicative competence.

Keywords TOEFL IBT • Communicative language testing concept • The adult colleges • Evaluation system reform

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN020

Analysis of Content Validity of Reading Comprehension in PRETCO-B

Presenter
Weihui Cui, Weili Cui
Langfang Yanjing Vocational and Technical College
Langfang, Hebei Province, China

Fang Zhao
Shijiazhuang University
Shijiazhuang, Hebei Province, China

Reading comprehension is usually the most important and difficult part in PRETCO-B. This paper makes a study on the content validity of reading comprehension in PRETCO-B on the basis of language testing theory. This study analyzes the content validity of reading comprehension in PRETCO-B. In detail, this paper makes the qualitative and quantitative analysis on reading comprehension from three aspects, which are the subjects, genres and readability, and puts forward corresponding solutions. This study aims to providing a more scientific and objective standard for reading comprehension in PRETCO-B, and providing some valuable research data for the future study in this area.

Keywords Language testing theory • Content validity • Reading comprehension • PRETCO-B

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN022

The Application of Descriptive Statistical Analysis of Foreign Language Teaching and Testing

Presenter
Wei Wu
Hebei University of Economics and Business
Shijiazhuang, Hebei Province, China
E-mail: laimerwei@126.com

In foreign language teaching tests, various tests continue, because the teaching is inseparable from the test. Conversely, the test cannot be separated and then analyzed the results of statistics and data. In the traditional way of language testing, teachers usually analyze the scores by calculating the mean and classifying point segments. Though easy to operate, the method of mean calculation is susceptible to extreme data and the classification of point segments cannot tell the distribution of scores when they are abnormally distributed, resulting in the unavailability of positive feedback on teaching from the test. Therefore, this paper proposes the use of Descriptive Statistical Analysis and the Combination of Measures of Central Tendency, Divergence Tendency and Standard Score, which will find out valuable information to enhance the teaching quality.

Keywords Descriptive statistical analysis • Central tendency • Divergence tendency • Standard score • Foreign language teaching

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN024

Error Analysis of English Writing in College English Placement Test of Hebei Normal University

Presenter
Fang Zhao
Shijiazhuang University
Shijiazhuang, Hebei Province, China
E-mail: jfcyu@163.com

Writing competence has been playing a significant role in evaluating Chinese English learners' proficiency, while writing performance of students is far from satisfactory according to the statistics declared officially from the examination center of CET-4 & CET-6. Research on writing in CET-4 displays that the average score of writing is far lower than the other items. Aims to foreign language writing teaching, the researchers prone to analyze the errors occurred in which with relevant theories. However, relatively few empirical studies have been conducted in investigating the errors in writing at each level.

Based on the error analysis of the first-year non-English major students' writings on computer-based College English Placement Test (CEPT) in Hebei Normal University, this dissertation attempts to find out the distribution and frequency of errors in writing at each level so as to help language teachers adopt appropriate teaching strategies and methods of error correction, then to promote their profession of writing teaching for English learners with the same proficiency in writing.

Keywords Error analysis • English writing • College English placement test

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN025

Fluency: A Critically Important Yet Undervalued Dimension of L2 Vocabulary Knowledge

Presenter
Junyu Cheng
School of Foreign Languages, Southeast University
Nanjing, China
E-mail: chjy@seu.edu.cn

Joshua Matthews
School of Education, University of Newcastle
Callaghan, Australia
E-mail: joshua.matthews@uon.edu.au

Background

Measurements of vocabulary knowledge are powerful indicators of language proficiency as they correlate with other language skills such as reading, writing, listening and speaking (Milton 2009; Stæhr 2008). It is evident to all L2 language educators that L2 vocabulary knowledge is an essential component of L2 competency. What may be less apparent, however, is that vocabulary knowledge is a construct which is most usefully considered as being comprised of a number of dimensions which act in concert to support a learner's general vocabulary competence. Among these knowledge dimensions, vocabulary fluency, or the ability to access and apply vocabulary knowledge under time constraints, is arguably the most important. This is especially the case in regard to the ability to apply existing L2 vocabulary knowledge in a communicatively competent manner.

Aims

This paper asserts that the ability to apply vocabulary knowledge in a fluent manner has been systemically undervalued in contemporary L2 educational contexts. To explore this assertion, two constructs of ESL vocabulary knowledge were measured among a cohort of ESL learners. One construct was operationalised as a fluency non-dependent form-meaning mapping task. The other was a fluency dependent word form recognition and production task.

Keywords L2 vocabulary knowledge • Vocabulary fluency • Vocabulary testing • High frequency vocabulary • Academic vocabulary

Sample

Tests were administered among a cohort of 113 intermediate Chinese tertiary level students (mean age = 18.5 years, SD = 0.74) studying within China.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Method

The two test types were administered to measure the high frequency (Nation 2001) and academic word knowledge of the participants (Coxhead 2000). The results from the tests provided data enabling comparison between the cohorts' collective *fluency dependent* and *fluency non-dependent word knowledge* for the two targeted word categories (high frequency and academic words).

Results

For the test format which did not depend on fluency, mean scores for high frequency vocabulary (M = 84.7 %, SD = 15.9) and academic vocabulary (M = 62.5 %, SD = 20.7) were significant higher than those for the corresponding fluency-dependent test format (high frequency vocabulary: M = 44.79 %, SD = 18.8, academic vocabulary: M = 28.6 %, SD = 15.9).

Conclusions

The cohorts' fluency-dependent knowledge of high frequency and academic words was significantly lower than knowledge of those same categories of words as measured by a fluency non-dependent test format. These results encourage an invigoration of teaching, testing and learning approaches which emphasize the development of vocabulary knowledge which can be applied in a fluent manner. Practical suggestions on improving the vocabulary fluency of ESL students are provided.

Correlation Analysis on Influencing Factors of Writing Development of English and Non-English Majors in Tibetan College

Presenter
Baina Xia
School of Foreign Languages, Xizang University for Nationalities
Xianyang, Shanxi Province, China
E-mail: xiabaina@hotmail.com

Aims

This study aims to research on the correlations between the five aspects in English writing, viz., content, organization, vocabulary, language use, and mechanics, and the three important factors that affect students' writing quality, viz., learners' English proficiency, writing strategy, and feedback (Ferris 1995; Mohan 1985; Paulus 1999; Santangelo et al. 2007; Sasaki 2000; Sexton et al. 1998; Wang and Dong 2010; Xiao 2008).

Keywords English writing • English proficiency • Writing strategy • Feedback • Content • Organization • Vocabulary • Language use • Mechanics

Methods

The researcher compared statistically 108 non-English major Tibetan college students' writing score in their first year of college. During this academic year, they were required to take the English writing course; the evaluating scheme of the tests and compositions stayed the same. A hundred and eighteen English compositions of eight students in this year were investigated and analyzed to reveal the development of the five aspects in their writing. Five of these eight students were interviewed in the period of their writing course so that more influencing factors can be discovered. By the end of their study in the first year, questionnaires were distributed to the 108 students.

Results

The data from their questionnaire and five aspects of their writing score were analyzed statistically. The analysis of the above-mentioned data show that: (1) English writing quality of these learners improved prominently after study in college for one year; (2) four of the five aspects in writing—content, organization, vocabulary, and language use-have shown improvement of different levels with exceptional but explainable changes, and the interviews can explain why the improvement is not persistent, while mechanics does not have a stable pattern of development; (3) in most cases, content, organization, vocabulary, and language use can be predicted by English proficiency and writing strategy, but cannot be

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

predicted by feedback; mechanics cannot be predicted by these three influencing factors.

Conclusions

In the end, implications of this study and suggestions for further studies are proposed. If positive factors can be reinforced and negative ones deleted, both English writing's learning and teaching can be more effective.

PROMS2014CN041

Measuring the Effectiveness of Evaluation Processes for Diversified Undergraduate Students with S-regulated Study—A Case Study in International School of Jinan University

Presenter
Yufan Liu
International School, Jinan University
Guangzhou, China
E-mail: tiuyufan@jnu.edu.cn

There are more and more universities in China accepting international students from various countries. These students with international background and highly diversified preferences in learning have presented challenges for educators in universities. Using the help of web-based learning platform and being guided in self-regulated study groups, students are now given much more alternatives in terms of learning and communicating. This paper first studies the difference in characteristics and study needs of undergraduate students. Then the effectiveness of web-based evaluation applied in self-regulated and heterogeneous teaching objectives is examined.

Keywords Web-based learning platform • Self-regulated study • Cooperative study • Economics course

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014CN047

An Rasch-Based Analysis on the Pragmatic Functions of the Negative Interrogative Sentence Pattern “ではないか”

Presenter
Hairu Yang
College of Foreign Studies, Jiaxing University

From such three perspectives as “the speaker has the doubt”, “the speaker has no doubt”, and “the speaker expresses his/her true opinion”, the writer makes a study on the pragmatic function of the sentence “ではないか”, based on the comparison of its pronunciation and tone as well as its semantic analysis. Different pronunciations and tones express different meanings, so Japanese language learners need to be careful in using the sentence.

Keywords Asking for confirmation • Doubt • Interrogation • Intention • Pragmatic functions • Euphemism

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014HK003

The Generalized Rasch Ipsative Model for No-preference Responses in Pairwise Comparisons

Presenter
Kuan-Yu Jin, Wen-Chung Wang
The Hong Kong Institute of Education

Background

In pairwise comparisons, respondents are requested to select one statement from a pair of statements. Several item response theory models have been developed for pairwise comparisons. In practice, respondents may choose an option of "no preference" (when this option is provided), or simply leave a blank when they have difficulty in making a choice. The no-preference responses (or blanks) are often treated as ignorable missing data and, in consequence, disregarded in the following analysis. Current analyses fail to consider the underlying mental process of generating no-preference responses.

Aims

A generalized Rasch ipsative model (GRIM) is developed to account for no-preference responses (or blanks) in pairwise comparisons, in which the probability of endorsing a no-preference option (or a blank) is determined by a personal threshold as well as the difference in utility between two given stimuli. The higher the personal threshold than the difference, the higher the probability of endorsing a no-preference option (or a blank) will be.

Samples

A survey of 303 students showing their preferences of six universities was selected for illustration. The dataset can be found in the Royal Statistical Society Website (http://www.blackwellpublishing.com/rss/Readmefiles/dittrich.htm)

Method

A brief simulation was conducted to examine parameter recovery of the GRIM. The computer program WinBUGS was used for parameter estimation. Afterwards, a survey of students' preferences of universities was analyzed by fitting the Bradley–Terry–Luce model (or equivalently the one-facet Rasch model) and the GRIM. Model data fit and parameter estimates were compared.

Results

As expected, the parameters were recovered fairly well, suggesting the GRIM is feasible. In the empirical example, although these two models yielded nearly identical estimates for item parameters, the GRIM provided additional information about the variations on the personal thresholds.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Conclusion

The proposed GRIM is feasible and useful in considering no-preference responses in pairwise comparisons.

PROMS2014HK010

Assessing the Usage of Hand Hygiene Observational Survey Tool

Presenter
Dr. Wai-fong Chan
Infection preventionist, Tung Wah Eastern Hospital, Hong Kong SAR
Visiting Lecturer, Tung Wah College, Hong Kong SAR
E-mail: wfchan@alumni.cuhk.net

Background

Hand hygiene is the single most important measure to prevent the infection spread in healthcare setting. To minimize the Hawthorne effect induced by the single observer in the hospital, a number of frontline staff was trained to join the hand hygiene observational survey in the future.

Aims

The aim of this study is to examine the assessment results of the training program for hand hygiene observers for program improvement.

Keywords Hand hygiene • Observer • Training

Sample

Fifteen Infection Control Link Persons or delegates who completed the training program and the assessment component.

Method

A classroom training program was conducted for potential hand hygiene observers in October 2013 to prepare them to observe the hand hygiene behavior of the healthcare workers on napkin changing procedure with a standardized survey tool. After the training program, the potential observers were requested to complete an assessment in order to validate their accuracy of utilizing the survey tool. The assessment included a paper presentation of eight scenarios of napkin change. Twenty-nine checkpoints were developed to check the accuracy of the usage of survey tool. Each checkpoint reviewed the documentation of hand hygiene indication (one or two) with corresponding observed behavior. Correctness/incorrectness of each checkpoint was recorded for individual potential observers. The data were analyzed by Winsteps dichotomous model 3.61.2.

Results

Data from 15 persons and 29 items were collected. No person got all the items correct. Person reliability ranged from 0.77 to 0.82. Six items belonged to minimum estimated measures. Item reliability ranged between 0.70 and 0.75 after removing the extreme items. The item measures spread from −3.95 to 6.19 while the person

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

measures lied between −2.83 and 5.41. There was a gap of 2.33 logits between person and item mean measures.

Conclusions

The potential observers grasped the basic principle of usage of the survey tool. Some complex scenarios are needed to be emphasized in subsequent coaching and training programs.

PROMS2014HK026

Using the Many-Facets Rasch Measurement to Improve the Scoring and Moderation Procedures in Writing Assessment

Presenter
Kinnie Kin Yee Chan
The Open University of Hong Kong

Background
Maintaining fairness and consistency in essay scoring both in the classroom and high-stakes assessment is crucial as the results of essay scoring will impact the development of students either directly or indirectly. Thus, teachers, students and parents are under considerable pressure in the education system in Hong Kong.

Aims
The Rasch model is applied to co-calibrate the scales for scoring of an Automated Essay Scoring (AES) system, the Lexile Analyzer, and human raters on essay writing.

Keywords English essay assessment • Human rating • Many-facets Rasch Measurement (MFRM)

Methods and sample
137 Hong Kong students responded to three essay writing prompts from the National Assessment of Educational Progress (NAEP) and essays were human raters scored by using the NAEP holistic essay marking rubrics covering the narrative, informative and persuasive genres. The four trained raters scored sets of the written responses of students. All the essays in this study were then scored by the AES engine. The study is on structuring data for the Many-facets Rasch Measurement (MFRM) analysis to adjust for prompt difficulty and rater severity and whole-instrument replication to perform an appropriate assessment of the precision of the writing assessment.

Results
The MFRM analysis reveals that the grades assigned by raters and machine show considerable overlap, with contrasts for some genres. Then, a post-scoring meeting was conducted with the four HK raters. They rescored some student essays and they were asked some questions to reflect how they would proceed to finalize, qualitatively, fair scores for the essays and how the results of essay scoring might affect students' development.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Conclusion

The results of the study will give more choices for teachers for the moderation of scoring in English essay assessment, and subsequent related learning and teaching strategies.

PROMS2014JP003

Effects of Judging Plans on Perceived Rater Strictness and Fit

Presenter
Jeffrey Durand
Junior Associate Professor, Tokai University, Japan

Background

This study investigates the effects of judging plans on tests involving rated observations, as is often done with speaking and writing exams. Judging plans, assignments of raters to examinees, link raters to each other as well when two or more raters are assigned to score the same examinee. The connections among raters create a network that is essential for Facets-based Rasch analysis. This kind of analysis provides estimates of examinee ability based on performance on specific items, but also taking into account rater strictness. According to Linacre (Judging Plans and Facets, http://www.rasch.org/rn3.htm), the only necessity is that the judging plan links all raters together in a network. The characteristics of the network, however, may impact test results.

Aims

This research looks at the how the network shape, i.e. the way raters are assigned together, affects test results when a rater is judging *in a random manner*. Network shape (chain or ring lattice), number of raters and network diameter, and number of rating partners (related to clustering) are investigated. The goal of this research is to provide guidelines that help reduce network effects on test scores.

Methods

This study used computer simulation with R statistical computing platform to generate data. Ability levels for 300 examinees were randomly generated. The simulated test contained six items. The first two used a common rating scale with varying difficulty, as did the latter four items. Rater strictness was also generated on each specified network. The measures and category boundaries were used to calculate the score probabilities of a particular item. A random uniform number between 0 and 1 was generated and used to assign scores. This was repeated to create 100 data sets with no rater randomness. For comparison, the process was repeated, but the strictness of one of the raters was adjusted by +2 for half of the ratings and −2 for the other half. In this way, randomness was added to the data. All data sets were analyzed via batch mode using Facets software. Average student ability and fit along with average rater strictness and fit were calculated.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Results

Results suggest that network shape and especially the location of the randomness affect student and rater measures. In addition, the randomness in one rater can make others ‘unfairly’ appear inconsistent as well. Network diameter and number of raters has little effect on measures and fit scores, though having more rating partners somewhat reduces the effects of randomness around the network.

Conclusions

In conclusion, when test administrators devise judging plans, network shape deserves consideration. More research is needed in this area, however. Frequent changing of rating partners may be more desirable than keeping raters together. Finally, for purposes of rater feedback, raters with poor fit scores may not in fact have been inconsistent.

PROMS2014JP005

Measuring Reading Speed Gains Using Many-Faceted Rasch Measurement

Presenter
Trevor A Holster
Fukuoka Women’s University
E-mail: trevholster@gmail.com
J. W Lake
Fukuoka Jogakuin University

Background

Reading quickly is essential if students are to cope with the quantities of reading material required for university study. Measuring the reading speed of students entering English for academic purposes (EAP) programs is problematic, however, because they may find authentic academic texts incomprehensible. Quinn et al. (2007) addressed this by providing 20 simplified texts of 550 words each, restricted to the first 1000 words of West’s (1953) General Service List, with the intention that words-per-minute scores of reading speed could be compared between texts and also between different points in time.

Aims

This study aimed to measure the difficulty of the texts provided by Quinn et al. (2007) using many-faceted Rasch measurement (MFRM) to confirm two hypotheses: (i) that the texts were of equal difficulty, and (ii) that raw words-per-minute scores provide the equal interval measures of ability required to compare gains in reading speed by different groups at different times.

Sample

Approximately 200 students enrolled in an academic English program for two semesters at a Japanese women's university provided the main sample, with approximately 50 male engineering students enrolled in an academic reading class for one semester at a Japanese public university providing supplementary anchoring data.

Method

Reading speed was measured at the beginning, middle, and end of each semester, with each student reading two or three texts per administration. The order of administration of texts was varied by class group, allowing measurement of the facets of student ability, text difficulty, and time. Data were analyzed using the *Facets* software package for MFRM (Linacre 1994).

Results

Two important findings were made. The range of difficulty of the 20 texts provided by Quinn et al. (2007) exceeded 2.5 logits, corresponding to about a 30 %

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

difference in raw reading speed. Secondly, reading speed measured in words-per-minute did not provide interval level measures of ability, but when these were converted to a logarithmic scale, they approximated interval level measures sufficiently for classroom purposes.

Conclusions

The 20 texts provided by Quinn et al. (2007) do not constitute parallel forms and are unsuitable for research purposes requiring additive comparison of scores derived from different subsets of the texts. Additionally, the use of raw words-per-minute scores is inappropriate for purposes requiring additive comparison of scores at different levels of ability and should be converted to a logarithmic scale if gains in reading speed over multiple administrations are to be compared.

PROMS2014JP019

The Discrepancy Between Accuracy and Confidence in Knowledge of English Verbs of Utterance

Presenter
Aaron Olaf Batty
A visiting lecturer at Keio University's Shonan Fujisawa Campus

Background

English verbs of utterance (e.g. "speak," "talk," "say," and "tell"), despite being basic vocabulary items, exhibit a wide range of uses by native speakers which are not predictable based on dictionary word meaning alone (e.g. "talk politics"). However, it is possible that as learners have more exposure to the language that they become more accurate and/or more confident with these special uses.

Aims

The present research investigates the discrepancy between accuracy, and confidence with regard to special uses of the utterance verbs "speak," "talk," "say," and "tell" among Japanese students of English at various levels of proficiency.

Keywords Rasch • English linguistics • Accuracy versus confidence

Sample

Japanese high school students (n = 22), university students (n = 140), and native speakers living in the USA, the UK, and Japan (n = 15; N = 177).

Method

A vocabulary test and vocabulary questionnaire was administered on paper to the non-native speakers of English during normal class time or online at their leisure. The questionnaire required participants to select the correct utterance verb to complete a sentence, and then indicate their degree of confidence for their answer. The test data were scaled in Facets (Linacre 2012) and compared to the questionnaire data via the method developed by Paek et al. (2008).

Results

Overall, respondents were underconfident of their knowledge of special uses of "say" and "tell," and generally overconfident of their knowledge of special uses of "speak" and "talk." The overall percentage of discrepancy observed was approximately 7 % underconfident for all the words and uses on the test, and the highest percentage of overconfident discrepancy was found to be approximately 10 % in the case of the verb "talk."

Conclusions

Overconfidence does not seem to be a serious problem with Japanese students of English with regard to these special uses of familiar verbs of utterance. Although

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

many of the items were quite difficult for the respondents, they were aware of their deficiencies. The findings regarding the verb with the highest degree of overconfidence—“talk”—may warrant explicit instruction on these special use cases, as students appear to be unaware of them, and seem to have acquired a sense of the verb that is lacking some of the core features of its meaning. Finally, the Paek et al. method was found to be relatively easy to apply to data of this kind, and the present author recommends it to researchers seeking to investigate issues of accuracy versus confidence.

PROMS2014JP021

Writing Assessment in University Entrance Examinations: The Case of One Japanese University

Presenter
Kristy King Takagi
Professor of University of Fukui
Fukui, Japan
E-mail: kjktakagi@hotmail.com

Background

Recent research includes evaluation of methods of writing assessment, prediction of academic success for Japanese students attending an international university in Japan, and evaluation of the Japanese national Center Examination.

Aims

The aim of the project was to analyze the performance of raters and the rating instrument used in assessing essays for university placement testing in a Japanese international university.

Keywords University testing • Essay rating • Placement testing

Methods

It was hypothesized that the raters, despite their advanced education and experience in rating essays, would not rate uniformly, and that the categories of the rating instrument would not be consistently used. In order to test this hypothesis, ratings given by 15 EAP instructors to 133 essays written by first-year Japanese students at an international university in Japan were analyzed. The FACETS computer program was used to evaluate rater and rating instrument performance.

Results

Results showed that the raters used the rating scale in quite different ways. For example, some raters had overfitting ratings; that is, they assessed the essays in limited and predictable ways. Other raters displayed another problem in that they were inconsistent and unpredictable. Only the ratings of two of the 15 raters fit the data well. As for the rating instrument, though a number of findings were positive, there were problems. Three of the nine score categories were not used at all, and the distances between four categories of scores were too small, revealing that each of these score categories was not distinctly defined.

Conclusions

In short, the results indicated that raters need ongoing training and feedback in order to improve their performance, and that essay rating instruments should be chosen with care. In addition, results suggest that employing additional methods of writing assessment could improve the reliability of placement decisions.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY003

Evaluating the Internet Addiction Scale for Chinese Students: The Application of Rasch Model

Presenter
Xiuxiu Qian
College of Foreign Studies, Jiaxing University

A.Y.M. Atiquil Islam
Institute of Graduate Studies (IGS), University of Malaya
E-mail: skyiium@yahoo.com

In the 21st century researchers recognize the huge potential of Internet in improving educational outcomes, and in promoting research among students and academics in higher education. However, the Internet was revealed to be addicted, especially by students. As such, the aim of this study is to validate the Internet addiction scale for evaluating students' addiction in using Internet for their learning purposes. To validate the Internet addiction scale, data achieved through a survey conducted with 264 students studying in four colleges (Foreign Studies, Business, Education, and Science) of a university in China. A five-point Likert scale that indicated degrees of agreement/disagreement was used to capture the students' views about the Internet addiction. The instrument reliability and validity were conducted by Rasch model applying Winsteps version 3.49. The findings of Rasch analyses discovered that (i) the items reliability was found to be at 0.98 (SD = 105.7), while the persons reliability was 0.82 (SD = 9.4); (ii) the items and persons separation were 7.39 and 2.14 respectively; (iii) all the items measured in the same direction (ptmea. corr. >0.38); (iv) all items showed good item fit and constructed a continuum of increasing intensity. The findings also demonstrated that the Rasch model is applicable to validate the Internet addiction scale and it fosters support for the internal consistency and unidimensionality.

Keywords Internet addiction • Rasch Model • Higher education

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY005

Predicting Students ICT Usage in Higher Education

Presenter
A.Y.M. Atiquil Islam
Institute of Graduate Studies (IGS), University of Malaya
Email: skyiium@yahoo.com; skyum2013@gmail.com

Background

As information and communication technology (ICT) is increasingly used in higher education, its integration to learning has had immense significance in fostering technology-based education among the students of a university. However, extensive research has been demonstrated that the ICT facilities were underutilized.

Aims

As a result, this study is to examine the underlying factors influencing the ICT usage among students in higher education. An extended Technology Acceptance Model (TAM) was used as the theoretical framework where it was predicted that the ICT's perceived ease of use and perceived usefulness, in addition to students' computer self-efficacy, would positively influence on students ICT usage in Higher education.

Keywords Information and communication technology • Extended technology acceptance model • Structural equation modeling • Higher education

Sample

A total of 250 students from seven faculties (Education, Information and Communication Technology, Engineering, Islamic Revealed Knowledge and Human Sciences, Economics and Management Sciences, Law, and Architecture and Environmental Design) were collected.

Method

The questionnaire's validity was conducted using factor analysis. The hypotheses were tested applying the Structural Equation Modeling.

Findings

The findings of this study showed that computer self-efficacy had a statistically significant direct effect on perceived usefulness and perceived ease of use. Subsequently, perceived ease of use had a statistically significant positive direct effect on use of ICT. Similarly, perceived usefulness had also significant direct effect on perceived ease of use. On the other hand, computer self-efficacy had a significant indirect effect on use mediated by only perceived ease of use. Moreover, computer self-efficacy also revealed a statistically significant indirect effect on use mediated by perceived usefulness and perceived ease of use. Eventually, perceived usefulness discovered a statistically significant indirect effect on use mediated by

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

perceived ease of use. The results also demonstrated that the extended TAM was found to be validated for predicting students ICT usage in higher education in terms of computer self-efficacy, perceived ease of use and perceived usefulness.

PROMS2014MY007

Examining Mobile Learning Acceptance Scale Using Rasch Model

Presenter
A.Y.M. Atiquil Islam
Institute of Graduate Studies (IGS), University of Malaya
E-mail: skyiium@yahoo.com

Nada Mansour F Aljuaid
Faculty of Education, Universiti Teknologi Malaysia

Mohammed Ali Rajab Alzahrani
Faculty of Computing, Universiti Teknologi Malaysia

Mobile learning has been advanced the learning processes for the learners in this technological era. However, literatures demonstrate that mobile learning is still under implementation, especially in higher education for teaching and learning. Thus, the purpose of this study is to develop and validate the mobile learning acceptance scale for examining students' intention to use mobile learning in higher education. A total of 140 students were selected from a comprehensive public university in Saudi Arabia using online survey. A set of questionnaire containing validated items from prior studies was put together and modified to suit the current study. A five-point Likert scale seeking the opinions of the respondents to the extent of their agreement/disagreement to the items was used. The questionnaires' reliability and validity were performed through a Rasch model applying Winsteps version 3.49. The results of Rasch analyses showed that (i) the items reliability was found to be at 0.90 (SD = 27.8), while the persons reliability was 0.93 (SD = 19.1); (ii) the items and persons separation were 2.93 and 3.66 respectively; (iii) all the items measured in the same direction (ptmea. corr. >0.46); (iv) the majority items revealed good item fit and constructed a continuum of increasing intensity. The results also indicated that the variance explained by the measures was 66.5 % which demonstrated that the items were able to endorse students' acceptance in using mobile learning in Saudi Arabia higher education for their learning purposes.

Keywords Mobile learning • Acceptance • Higher education • Rasch Model

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY016

The Application of Rasch in the Validation of Corporate Citizenship Scale

Presenter
Kamala Vainy Kanapathi Pillai (Ph.D)
School of Business, Curtin University
Sarawak, Malaysia
E-mail: kamala.pillai@curtin.edu.my

The paper articulates the application of Rasch measurement in corporate citizenship research. With burgeoning expectation for greater corporate responsibility, studies have found that many companies continue to resort to green washing tactics in order to cope with growing pressures. In the absence of systematic adoption, the concept of corporate citizenship may remain rhetoric. The study is aimed at determining the fundamental attributes that facilitate the internalization of corporate citizenship within companies. The study had to address two main challenges: Firstly, the small sample size expected as it was based on primary data collection explicitly seeking the views of managers practicing corporate citizenship; and secondly, the lack of prior systematic research available based on a multi-disciplinary approach. Rasch was applied to establish a psychometrically sound scale. A pilot test with 30 companies was followed with a larger sample set. A total of 634 companies listed on the Malaysian Exchange were surveyed through online survey, out of which 100 companies responded. The instruments' reliability and validity were conducted using Winsteps version 3.49. The results of Rasch modelling analysis indicated the items measured reliability ($r = 0.93$), and persons measured reliability ($r = 0.97$). In addition, both item separation (3.51) and person separation (5.27) were found to be statistically significant. Further, all items measured in the same direction (ptmea. corr >0.30) and valid items showed good item fit and constructed a continuum of increasing intensity. This study's significance stands on its contribution towards the application of Rasch by future researchers in diverse educational and industrial settings in developing and validating sound scales.

Keywords Rasch Model • Corporate citizenship • Corporate conduct • Stakeholder collaboration

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY017

Psychometric Rasch Analysis of Educational Persistence Test

Presenter
Monsurat Olusola Mosaku
Department of Educational Foundation, Faculty of Education,
Universiti Teknologi Malaysia
E-mail: mabaqo@ymail.com

Mohamed Najib Abdul Ghafar
Department of Educational Foundation, Faculty of Education,
Universiti Teknologi Malaysia
E-mail: p_najib@utm.my

Accountability, competitive edge in the global economy and employees increased expectations to mention a few, face higher education stakeholders in the twenty-first century. In order to surmount these and other issues, psychological variables comes into foreplay and their relevance is of great importance as a large body of research has examined the link between various psychological variables and academic attainment in higher education. One of such psychological variable which is of primary importance in this research is educational persistence. Educational persistence is operationally defined as analysis of the learning problem from the mind frame of the individual's goal orientation, advancement of synthesized (attention) and alternate strategies (self-regulation) to approach and solve the learning problem, while maintaining an intrinsic system of thoughts (disposition to persevere); all with the aim to aid students' successful degree completion. Despite the importance, effect size and "hit rates" of psychological variables based on literature review, a combination of non-intellective psychological predictors namely goal orientation; attention; disposition to persevere and self-regulatory strategies has however not been investigated. Furthermore, within cognitive psychology, neuroscience is gaining prominence but its caveat still exists within academic attainment most especially in higher education learning and warrants for its inclusion. As such, it is imperative to originally develop and validate an instrument composed of a combination of non-intellective psychological predictors for a multicultural higher education context. A quantitative cross-sectional, exploratory survey research method was implemented within a representative public university with a sample size of 428 students. Educational persistence development and validation process consisted item generation, slimming and refinement; and Rasch Measurement Model analysis using Winsteps version 3.68.2 assessed the psychometric properties of dimensionality; item polarity; item and person fitness; and response category function. The first stage of the Rasch Measurement Model analysis portrayed twenty-two misfitting items and necessitated the collapse of the five Likert-like

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

rating scale to a four Likert-like rating scale. Though the questionnaire, at the second stage of the Rasch analysis, had excellent psychometric indices, inferring that the items possess qualities of a good measurement instrument, the dimensionality is poor and more targeted items are required to depict high persistent levels students.

Keywords Educational persistence • Higher education • Rasch analysis • Test development

PROMS2014MY018

Malay Owner Managers of SMEs: The Typology

Presenter
Rohani Mohd
Malaysian Academy of SMEs and Entrepreneurship Development & Faculty of Business Management, UniversitiTeknologi MARA, Selangor, Malaysia

KhulidaKirana Yahya
Universiti Utara Malaysia, Kedah, Malaysia
BadrulHisham Kamaruddin, Mazzini Muda and Anizah Zainuddin
Universiti Teknologi MARA, Selangor, Malaysia

This study identifies different classification of Malay owner managers based on the psychological and behavioral factors. The survey comprised sample of 162 small scale SMEs' owners in manufacturing industry, in Malaysia. Rasch Measurement Model was employed for that purpose. The findings indicated that psychological (personal values and self efficacy motivation) and behavioral (entrepreneurial orientations) factors were able to differentiate owner managers into five classification; Exemplar, Competence, Mediocre, Survivor and Poor. The interesting findings serve as a reminder to the Malaysian government to focus on both the psychological and behavioral aspects of owner managers in an effort to improve Malay SMEs businesses; that is to develop programs that could improve their confidence to work under pressure, to make decision under uncertainty, at the same time to make them believe that to be unique and the best were not wrong in Islam; to have courage and self discipline were the key values to business success.

Keywords Typology • Personal values • Self-efficacy • Entrepreneurial orientations • Malay SMEs

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY020

Development and Validation of E-portfolio Adoption Scale in Higher Education Applying Rasch Model

Presenter
Mohammed Ali Rajab Alzahrani
Faculty of Computing, Universiti Teknologi Malaysia

A.Y.M. Atiquil Islam
Institute of Graduate Studies (IGS), University of Malaya
E-mail: skyiium@yahoo.com

Assoc. Prof. Dr. Suhaimi bin Ibrahim
Advanced Informatics School, Universiti Teknologi Malaysia

In its effort to foster a stimulating teaching and learning environment, Saudi Arabian tertiary education has provided an e-portfolio system for educators and learners. However, no research has been conducted to evaluate the adoption of e-portfolio system within its context. As such, the aim of this study is to develop and validate the E-portfolio adoption scale for assessing the lecturers' adoption in using e-portfolio system for their teaching and learning in tertiary education. A total of 222 lecturers from eleven colleges (Education, Sciences, Medicine, Computers and Information Systems, Administration and Financial Sciences, Pharmacy, Arts, Engineering, Community and Continuous Education, Health Sciences, and Islamic Law) were selected through online survey of a comprehensive public university in Saudi Arabia. A seven-point Likert scale seeking the opinions of the respondents to the extent of their agreement/disagreement to the items was used. The instrument's reliability and validity were established through a Rasch Model using Winsteps version 3.49. The findings of Rasch model depicted that (i) the items reliability was found to be at 0.81 (SD = 31.2), while the persons reliability was 0.95 (SD = 47.3); (ii) the items and persons separation were 2.04 and 4.57 respectively; (iii) all the items measured in the same direction (ptmea. corr. >0.45); (iv) the most items represented good item fit and constructed a continuum of increasing intensity. The results also confirmed that the e-portfolio adoption scale was found to be validated to assess the lecturers' adoption and successful integration in using it for their teaching and learning in higher education. Moreover, the Rasch model was applicable to develop and validate the scale to generalize the idea of e-portfolio system implementation in higher education which could be applied by the future researchers in the diverse context of education.

Keywords E-portfolio • Higher education • Rasch Model • Adoption

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Assessment Literacy: What Do Pre-service Teachers Know About Assessment?

Presenter
Suah See Ling, Ph.D
Institute of Teacher Education Penang Campus
E-mail: seeling.suah@gmail.com

Assessment literacy in this study can be understood in terms of knowledge about test construction, types of assessment, use of assessment and grading & scoring. A 34-item Teacher Assessment Literacy Inventory (TALI) was completed by 401 pre-service teachers. The main purpose of this study was to examine the level of assessment literacy of pre-service teachers using Rasch Model analysis. TALI is a newly developed instrument to determine the level of assessment literacy of school teachers in Malaysia. The analysis was conducted with WINSTEPS 3.68. Item calibration of the 34 items of TALI was estimated, in order to identify the assessment literacy of pre-service teachers. To get a better view of the assessment literacy of pre-service teachers, the mean logit of the items belonging to a category was calculated. The 4 categories of assessment literacy were then ranked in order of difficulty based on mean logits. The distribution of mean logits ranged from −0.42 logit to 0.49 logit. This study found that the pre-service teachers had higher knowledge about "Types of Assessment" (−0.42 logit) whereas they had less knowledge in "Grading & Scoring" (0.49 logit). The mean ability estimates of the pre-service teachers (0.32 logit) were greater than the mean difficulty estimates of the items (0.00 logit). This suggested that the items were easy for the ability of the pre-service teachers. A total of 250 (62.34 %) pre-service teachers have the ability estimates above 0.00 logit. This data indicated that the level of pre-service teachers' assessment literacy was satisfactory. The findings provided the Malaysia's Ministry of Education with useful information for the training of school teachers in assessment and measurement.

Keywords Assessment literacy • Pre-service teachers • Rasch Model

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY029

Malaysian and Australian Students' Views on Diversity: A Differential Item Functioning Analysis

Presenter
Ssekamanya Siraje Abdallah
Institute of Education, International Islamic University Malaysia
E-mail: Siraje@iium.edu.my

Jac Brown
Department of Psychology, Macquarie University
E-mail: jac.brown@mq.edu.au

Jeanna Sutton
School of Psychology, University of Western Sydney
E-mail: j.sutton@uws.edu.au

Noor Lide Abu Kassim
Faculty of Dentistry, International Islamic University Malaysia
E-mail: NoorLide@iium.edu.my

Kamal J Badrasawi
Faculty of Dentistry, University of Malaya
E-mail: kamalbadrasawi@gmail.com

Background

With global travel and access to media from around the world, people are increasingly confronted with similarities and differences between themselves and others that may encourage building bridges as well as walls. Often reality is distorted fuelling misunderstandings and resentments which sometimes lead to devastating consequences. For example, much of the distrust and violence that is seen in the world is in part, fuelled by perceived differences. Thus, the concept of diversity and people's attitudes towards it may be a crucial predictor of conflict.

Objective

This paper aims to explore the differences and similarities between Malaysian and Australian university students on factors related to their attitudes towards diversity, including level of aggression, nationalism, socialization, and patriotism.

Keywords Diversity • Differential item functioning • Patriotism • Nationalism • Socialization • Emotion • Rasch Measurement Model

Method

A survey was administered to students at the International Islamic University of Malaysia (n = 398) and students at the University of Western Sydney (n = 239). The

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Rasch Measurement Model was utilized to investigate the differential item functioning (DIF) across the two nationalities.

Results

The analysis showed significant and substantial differences between Australian and Malaysian students on many items related to the different aspects of diversity. DIF contrast favoured Australian students on some items and vice-versa. Results of the DIF analysis support a general view of the two groups. More interestingly, the DIF analysis seems to point to a changing trend among young Malaysians.

Conclusion

To explore the factors that relate to attitudes towards diversity is important to understand how these attitudes are formed and how they evolve.

PROMS2014MY033

Validation of the Teachers' Performance Scale: The Application of Rasch Model

Presenter
Hasina Banu Shirin
Institute of Education, International Islamic University, Malaysia

A.Y.M. Atiquil Islam
Institute of Graduate Studies (IGS), University of Malaya, Malaysia

Xiuxiu Qian
College of Foreign Studies, Jiaxing University, China

Mohammad Serazul Islam
School of Business, University Kuala Lumpur, Malaysia

Teachers' performance depends on the retrospective of a teacher's expertise, experience, leadership quality, duty and responsibility. In other word, teachers' performance may be defined like other professionals i.e. teachers are required to determine what is vital for them to know and the process of demonstrating their existing knowledge and proficiencies through their vigorous enactment. Thus, teachers must need to have the potentiality to inculcate their knowledge, aptitudes and attitudes among learners. However, researchers have revealed that teachers are having lack of teaching methodological knowledge and leadership skills to educate secondary school students those will be the future role players for developing countries as well as nations in this globe. Likewise, recently some fluctuations and complications have been noticed in roles and responsibilities of leadership among secondary school teachers in Bangladesh. Hence, Bangladesh is one of the under developing countries and its present study puts in place with the hope of ensuring more efficient secondary school teachers who are able to educate students although, no research has been conducted to examine the teachers' performance within its context. As such, the purpose of this study is to validate the teachers' performance scale for assessing their abilities in educating learners. A total of 307 students from three secondary schools were selected using stratified random sampling technique to evaluate teachers' performance. A set of questionnaire containing validated items from prior studies was put together and modified to suit the current study. A seven-point Likert scale seeking the opinions of the respondents to the extent of their agreement/disagreement to the items was used. The data analysis was performed applying Rasch model using Winsteps version 3.49. The results demonstrated that (i) items and persons measured reliability (r = 0.99, and r = 0.75, respectively); (ii) the majority of items measured in the same direction (ptmea. corr. >0.30); (iii) most items showed good item fit and construct a continuum of

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

increasing intensity. The findings also revealed that the variance explained by the measures was 86.5 % which showed that the items were able to endorse the secondary school teachers' performance in teaching.

Keywords Teachers' performance • Secondary schools • Rasch Model

PROMS2014MY036

Reliability and Validity of the Malay Version of the SSIS Instrument for Student Aged 8- to 12-Years Old in Assessing Social Skills and Problem Behavior

Presenter
Zuraini Mat Issa, Ph.D Candidate
Department of Foodservice, Faculty of Hotel and Tourism Management, UniversitiTeknologi
E-mail: zurainim@salam.uitm.edu.my
Wan Abdul Manan Wan Muda
Professor, School of Health Sciences, Universiti Sains Malaysia
E-mail: wanmanan@usm.my

The Social Skills Improvement System (SSIS) Rating Scales for students can be used to evaluate student's social skills and problem behavior. The purpose of this study was to validate and examine the reliability of the Malay version of the SSIS instrument for student aged 8- to 12-year old using the item analysis method. The 75-items self-administered translated SSIS instrument was introduced to 188 students from two conveniently selected primary schools in Malaysia. The polytomous data were analyzed using the Winstep version 3.80.1 which applied Rasch measurement model based on Item Response Theory (IRT) Models. Item reliability index was used in examining the instrument reliability while fit statistics which include the point-measure correlation (PTMEA Corr) index and MNSQ values and unidimensionality were examined for instrument construct validity. The results showed that all SSIS subdomains have the reliability values of Cronbach's alpha >0.80 (0.92–0.98) and separation indexes >2 (3.38–7.40) with positive PTMEA values for all items. The Infit and Outfit MNSQ ranged 0.5–1.5 were used for the purpose of reviewing and retaining items. The findings also showed that 73 were fit items (MNSQ 0.5–1.50, none of the items were overfit and 2 items were misfits. Thus, only two items out of 75 items could be considered for removal. In conclusion, the Rasch measurement model could be used to produce empirical evidence of validity and reliability of the translated instrument. Hence, the Malay version of the SSIS instrument for student aged 8- to 12-year old could then be used to further assessing student's social skills and problem behavior.

Keywords SSIS rating scales • Primary school students • Rasch Measurement Model • Validity • Reliability

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014MY037

The Differential Contribution of Microsystems Factors to Boys' Behaviours in School and Outside School

Presenter
Jamilah Jaafar
Aminuddin Baki Institute, Ministry of Education Malaysia
Pahang, Malaysia
E-mail: jem@iab.edu.my

Background

Boys have been reported to have lower academic achievement and engagement in school. In many countries, boys are more likely than girls to dropout from school, repeat a grade, and have higher rates of suspension and expulsion. Recent data from a number of developed countries indicated lower rates of postsecondary enrolment and graduation among boys. Studies have also found that boys are more likely to be involved in anti-social behaviours such as drugs, alcohol, smoking, vandalism, illegal motorcycle racing, fighting, crimes and illicit sex. An increasing number of boys in government schools in Malaysia are showing negative attitudes in schools such as laziness, lack of motivation, and irresponsibility towards learning and school work.

Objective

Thus, the aim of this study was to examine the influence of microsystems factors, namely perceived parental support, teacher effectiveness and peer influence on three types of boys' behaviour in and outside schools; that is risk behaviour, positive school behaviour and negative classroom behaviour.

Keywords Structural equation modelling • Gender • Boys

Method

A survey was administered to 1309 boys randomly selected from 25 schools in Malaysia. Using Bronfenbrenner's Theory of Ecological Systems, the three microsystem factors were postulated to significantly positively influence positive behaviour, while negatively influencing risk behaviour and classroom behaviour. Confirmatory Factor Analysis (CFA) was employed to establish the reliability and validity of observed and latent variables, followed by Structural Model (SM) analysis to determine the causal factors.

Results

Findings of the CFA showed that microsystems factors achieved the divergent validity. Risk behaviours and positive behaviours were lowly correlated at −0.28. As expected, positive behaviours in school and negative classroom behaviours were moderately correlated at −0.58. In the Structural Measurement approach teachers'

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

effectiveness emerged as the most influential factor determining the boys' behaviours inside and outside the classroom although the correlations were at a moderate level. Peers' role emerged as being more important than parents in all behaviours measured in the study.

Conclusion

The entire model has achieved a good fit to the data suggesting that the micro-systems factors namely, perceived parental support, teacher effectiveness and peer influence, play a significant role in shaping boys' behaviours.

PROMS2014MY038

The Readiness of Secondary Schools Teachers in Palestine-Nablus to Adopt E-learning

Presenter
Fuad A.A. Trayek
Personalized Education Research Group, Faculty of Education, Universiti Kabangsaan Malaysia
E-mail: Fuad2004_a@hotmail.com

Tunku Badariah Tunku Ahmad
Institute of Education (INSTED), International Islamic University Malaysia
E-mail: Tunku26@hotmail.com

Rosseni Din
Personalized Education Research Group, Faculty of Education, Universiti Kabangsaan Malaysia
Email: rosseni@yahoo.com

Mohammed AM Dwikat
College of Engineering & Information Technology, An Najah National University Nablus, Palestine
E-mail: dwikatmo@najah.edu

Background

Electronic learning (e-learning) is an instructional method that delivers learning content electronically whether through synchronous or asynchronous online modes, or through multimedia platforms such as CD-ROMs or DVDs. In Israeli-occupied Palestine, e-learning is a viable means of offering continuous education to Palestinians, especially in war-stricken zones such as Nablus.

Objective

This study explored predictors of the readiness level of the secondary school teachers in Nablus to adopt e-learning in three important aspects: technological readiness, psychological readiness and equipment readiness.

Method

A total of 475 teachers (236 males and 239 females) sampled from 24 secondary schools in Nablus participated in the survey that employed a 23-item self constructed questionnaire measuring technological, psychological, and equipment readiness on a 5-point Likert scale. Multiple regression analysis was employed to determine the predictors of technological readiness while the Rasch Measurement Model was utilized as to investigate the differential item functioning (DIF) across

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

the independent variables of sex, age, highest academic qualification, teaching experience, access to a computer lab and availability of internet at school.

Results

Significant regression equations were found for all of the three aspects of readiness to adopt e-learning (i.e., technological readiness, psychological readiness, and equipment readiness. Sex, age, and highest academic qualification were significant predictors of technical and psychological readiness, while sex, access to a computer lab and the internet were significant predictors of equipment readiness. Results of the DIF analysis support a general view of the various independent variables. More interestingly, the DIF analysis seems to point to a changing trend, whereby younger teachers found most items easier as compared to older teachers.

Conclusion

The findings have important implications on what the Palestinian Ministry of Education and Nablus school principals need to put in place in order to better prepare their teachers and schools for e-learning.

PROMS2014SD001

Measuring Adolescent Perceptions of the Physical School Environment—An Analysis of the Psychometric Properties of a Scale Using Swedish Adolescent Data

Presenter
Daniel Bergh, Ph.D
Centre for Research on Child and Adolescent Mental Health, Karlstad University
SE-651 88 Karlstad, Sweden
E-mail: daniel.bergh@kau.se

Background
Adolescents spend a considerable amount of their time in the school environment. Most adolescents are also subjected to compulsory school attendance, implying that they have to deal with the environment on a daily basis. In health research adolescent perceptions about the school environment are often linked to mental and psychosomatic health. However, measurements seem to be focused on psychosocial or psychological aspects of the school environment more often than physical.

Aims
The purpose of the present study is to examine the psychometric properties of a scale of Adolescent Perceptions of the Physical School Environment by means of the Rasch model for ordered response categories.

Sample
The analysis is based on the survey Young in Värmland which is a paper-and-pencil based survey, conducted recurrently since 1988 targeting all adolescent in school year 9 residing the county of Värmland, Sweden. So far, more than 20,000 individuals have participated in the survey. In the analysis presented here, five items based on adolescents' perceptions of the physical school environment were subjected to analysis using RUMM2030, in total about 22,000 individuals.

Methods
A scale consisting of five polytomous items is analyzed by means of the polytomous Rasch model. General fit statistics as well as their graphical representations (ICC) are used to evaluate if the scale fit the Rasch model. A particular focus is also directed towards possible Differential Item Functioning (DIF) across sex.

Results
At a general level of analysis the scale subjected to analysis seems to fit the Rasch model fairly well, with good separation of the individuals, and showing no reversed item thresholds, i.e. the response categories work properly and as expected. Also, at a finer level of analysis focusing on DIF, the scale works fairly well, but with exceptions important in order to understand differences between boys and girls.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Conclusions

Although the scale fits the Rasch model fairly well, there is room for improvements. In particular the precision of measurement may be increased by improving the targeting through inclusion of additional items of appropriate severity.

PROMS2014SD002

Sample Size and Chi-Squared Statistics for Test of Fit—A Comparison Between a Random Sample Approach and a Chi-Square Value Adjustment Method

Presenter
Daniel Bergh, Ph.D
Centre for Research on Child and Adolescent Mental Health, Karlstad University
SE-651 88 Karlstad, Sweden
E-mail: daniel.bergh@kau.se

Background

Significance tests are commonly sensitive to sample size, and Chi-Squared statistics is not an exception. Nevertheless, Chi-Squared statistics are commonly used for test of fit of measurement models. Thus, for analysts working with very large (or very small) sample sizes this may require particular attention. However, several different approaches to handle a large sample size in test of fit analysis have been developed. Thus, one strategy may be to adjust the fit statistic to correspond to an equivalent sample of different size. This strategy has been implemented in the RUMM2030 software. Another strategy may be to adopt a random sample approach.

Aims

The RUMM2030 Chi-Square value adjustment facility has been available for a long time, but still there seems to a lack of studies describing the empirical consequences of adjusting a sample to a smaller effective sample in the statistical analysis of fit. Alternatively a random sample approach could be adopted in order to handle the large sample size problem. The purpose of this study was to analyze and compare these two strategies as test of fit approximations, using Swedish adolescent data.

Sample

The analysis is based on the survey Young in Värmland which is a paper-and-pencil based survey conducted recurrently since 1988 targeting all adolescent in school year 9 residing the county of Värmland, Sweden. So far, more than 20,000 individuals have participated in the survey. In the analysis presented here, seven items based on the adolescents, experiences of the psychosocial school environment were subjected to analysis, in total 21,088 individuals.

Methods

For the purposes of this study, the original sample size was adjusted to several different effective samples using the RUMM2030 adjustment function, in the test of fit analysis. In addition, 10 random samples for each sample size were drawn from the original sample, and averaged Chi-Square values calculated. The Chi-Square values obtained using the two strategies were compared.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

Results

Given the original sample of 21,088, adjusting to samples of 5,000 or larger, the RUMM2030 adjustment facility work as well as a random sample approach. In contrast, when adjusting to lower samples the adjustment function is less effective in approximating the Chi-Square value for an actual random sample of the relevant size. Hence, fit is exaggerated and misfit under estimated using the adjustment function.

Conclusion

Even though the inferences based on p-values may be the same despite big Chi-Square value differences between the two approaches, the danger of using fit statistics mechanically cannot be enough stressed.

PROMS2014SG005

The Stability of IRT Item-Parameters Estimation Using Rasch Model

Presenter
Mr Soo Kia Yong
An Examinations Development Officer in the Institute of Technical Education (ITE), Singapore
E-mail: cve_sky@yahoo.com.sg

Background

The classical test theory (CTT) and item response theory (IRT) are two competing frameworks in educational and psychological measurement (Hambleton and Jones 1993). Though CTT is conventionally used in test appraisal, it is often criticized for its two main limitations of “sample dependence” and of “test dependence” (Hambleton et al. 1991). On the other hand, IRT which has been heralded as “one of the most important methodological advances in psychological measurement in the past half century” (McKinley and Mills 1989), is believed to be superior for estimating item and person statistics compared to those obtained with the CTT framework (Acton 2003; Hambleton and Jones 1993; van der Linden 1986). Although IRT offers many benefits, the length of a test and the number of examinees needed for proper estimation of the item parameters are difficult to be determined (Hambleton 1989).

Most applications of IRT are found in large-scale testing situations (Sireci 1992). This is because bigger samples and longer tests are needed to provide accurate estimates, especially when both item and ability parameters are being estimated and when more complex IRT models are being fitted to the data (Hambleton 1989; Hulin et al. 1982). Unfortunately, many school-based assessments are administered to relatively small samples of examinees. Therefore, a critical question on the minds of educators, measurement specialists and test constructors is whether IRT models are able to obtain stable and accurate parameter estimates for small sample sizes.

Aims

The main objective(s) or the goal to achieve (where appropriate) and the relevant hypothesis or hypotheses tested (if there is).

This study was aimed at gathering empirical evidences of the psychometric characteristics of test-items on an end-of-module online assessment in Analogue Electronics administered to ITE Year 1 students taking the common module in National ITE Certificate (Nitec) in Electronics.

The objectives are:

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

(1) to investigate the effect of various sample sizes (25, 50, 100, 150, 200, 250, 300, 350, 400, 450 and 500) on the stability of item difficulty parameter estimation using real test data.
(2) to examine the effect of various sample sizes (25, 50, 100, 150, 200, 250, 300, 350, 400, 450 and 500) on the stability of test information function at different ability levels using real test data.

In essence, this study aims to find out whether the item difficulty parameter estimation and test information function at different ability levels are stable if small sample size is used.

Keywords Item Response Theory • Rasch analysis • Item difficulty parameter • Sample size • Test information function

Sample

The data source came from the end-of-module online assessment for a common module (Analogue Electronics) in Nitec in Electronics. Analogue Electronics is one of the core modules for students taking the courses in Nitec in Electronics. It is a foundation module which enables students to interpret, construct, test and troubleshoot analogue electronic circuits upon completion of the module.

The participants for this study were the ITE Year 1 students. These participants were post-secondary students with an average age of 17- to 19-year-old. The duration of the assessment was 1 h 15 min comprised of 30 multiple-choice items with 4 options. The students were assessed on six technical knowledge factors (Transformer, Power Supply, Transistor Amplifier, Operational Amplifier, Oscillator and Filter) and across three taxonomic levels (Recall, Comprehension and Application). Each correct item was awarded 1 mark and there was no penalty for wrong answer. A total sample of 500 participants was used in this study.

Method

The research design, method(s) and procedures (when necessary) and in the case of Rasch-based study, the relevant Rasch-based or IRT-based software needs specifying

The assessment was administered online to 500 examinees through an online portal called e-tutor. Raw data were collected and the item difficulty parameters were obtained and examined empirically for the 30 multiple-choice items. The computer software WINSTEPS was used for the Rasch calibration. The item difficulty parameters for the 30 items which were administered to the entire group of 500 examinees represented the population item difficulty parameters.

Smaller sample sizes were obtained through random sampling from the 500 examinees. The random samples were selected using MS Excel RAND function to obtain the require sample sizes. In addition, the RAND function is replicated ten times for every smaller sample sizes. The item difficulty parameters for the smaller sample sizes were calibrated and the average values represented the sample item difficulty parameters.

The current study borrowed the root mean square loss (RMSL) index to judge the accuracy of item difficulty parameters estimated. RMSL referred to the differences between the estimated and true parameters as an estimation loss (Han et al. 1997).

Similarly, test information function at different ability levels using population and sample item difficulty parameters were obtained. Hambleton and Swaminathan (1985) considered the test information function as the IRT analog of test score reliability. The index of relative efficiency, RE(θ), was computed for ability levels (θ) range -3 to 3.

Results

The item difficulty parameter appeared to be slightly sensitive to fluctuations with different sample sizes. The RMSL values obtained from various sample sizes were between 0.00 and 0.14. According to Han et al. (1997), RMSL values less than 0.6 were considered small. Similarly, the test information functions at different ability levels were comparable. The indices of relative efficiency computed were between 0.97 and 1.04.

Conclusions

The findings of this study show that the item difficulty parameter estimation and test information function at different ability levels are stable for different sample sizes when real test data undergone Rasch analysis.

With "data to drive instruction" in education circles lately, these findings have practical implications to the teachers who are in the journey of test appraising. Teachers will find the Rasch concepts (e.g., item-free ability estimates) and the applications useful in understanding students' testing behaviours. Besides, with technology advances which make computation and analysis of the assessment easier and efficient, teachers should gradually educate themselves in Rasch concepts which have better application potentials such as item banking and adaptive testing.

PROMS2014TW001

Rasch Analysis of Evaluation on Employees' Ability in Physical Performance Using SMARTC Physical Machine

Presenter
Hing-Man Wu, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
E-mail: smile@mail.chimei.org.tw

Background

Exercise and sport are very important for people, especially for those who are in a manner of geriatric (>65 years), middle age of sub-healthy adult (>40 years), and those with neurological disorder. However, no reliable objective measurement can be used for them.

Purpose

To evaluate person's performance in physical conditions using Rasch analysis and a SMARTC training machine.

Methods

A total of 57 workers in a hospital were recruited to assess their physical performances (e.g, movement momentum on watts usage, velocity, range of motion, and repeated motion times in 5 min) under their most comfortable conditions to administer the physical therapeutic machine. All data were transformed to be adaptive to Rasch analysis. A standardized score for each item was transformed as a rating score in a range between 0 and 5, depending on the standard deviation of items. An Excel-VBA module developed by authors was used to evaluate workers' ability in physical motion momentum. Ferguson (1949) delta coefficient for discriminating persons and Gini (1912) coefficient for measuring persons' inequality were applied.

Results

One dimension for the test was confirmed with a person reliability of 0. 886. Wright map was plotted in Excel showing all the item infit mean square errors near 1.0. Three clusters were obviously grouped for the study examines. A colorful Kidmap shown in Excel was more acceptable to readers than the traditional monochrome Kidmap. Ferguson delta and Gini coefficient are 0.53 and 0.39, respectively.

Conclusion

Values collected from a physical machine is essential for persons who want to periodically compare their own performance in the areas of physical agility,

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

balance, coordination, flexibility, strength, endurance and cardiopulmonary function. How to encourage persons, especially for middle age sub-healthy adult, to do exercise every day is required in order to achieve optimal harmony among mental, spiritual and physical conditions.

PROMS2014TW002

Comparisons of the Visibility of Portal Websites and the Key Words Used for Hospitals in Taiwan

Presenter
Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
E-mail: smile@mail.chimei.org.tw

Background
With the popularity of the Internet in recent years, hospital accreditation in Taiwan specifies that a portal website on health education be made available to the public.

Aims
There is a need to study (survey and rank) the visibility of portal websites for hospitals in Taiwan.

Keywords Portal website • Visibility • Search engine • World wide web • Rasch analysis

Sample
Five hundred thirteen hospital portal websites were specified in a sample.

Methods
Seventy-five keywords (as items) and 273 hospitals (as examinees) were fitted to the Rasch model (1960) after a serial Rasch analysis was performed. The extremely worldwide popular search engine, Google, incorporated with a computer program to facilitate searching efficiency was used in the present study. After >54,891 web page searches, a Likert-type 0–9 score was analyzed to estimate the visibility of each hospital portal website.

Results
There were 75 items and 275 hospitals that fit to the Rasch model expectation. The most difficult items to search on hospital websites were types of registered clinical professionals, followed by hospital beds and departments. The differences in visibility were substantially significant ($p < 0.05$) in paired comparisons between types of hospitals based on the following hierarchy using ANOVA one-way variance analysis: medical centers>region hospitals>local hospitals.

Conclusion
We suggest that hospital managers focus not only on website content, but on information architecture and meta-data to increase the visibility of a hospital portal website. The government section in charge of healthcare could develop an official website search engine to facilitate hospital accreditation based wholly on assessing the performance of website visibility.

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014TW008

Using Rasch Simulation Data to Verify Whether Ferguson's Delta Coefficient Can Report Students' Abilities Are Equal in a Class

Presenter
Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
E-mail: smile@mail.chimei.org.tw

Background
Many teachers are concerned about whether students' abilities are equal. The more equal students' abilities are, the more willing many teachers are to teach the class. A coefficient is required to compare the degree of equality between students' academic abilities. Ferguson's Delta (1949) an index of discrimination measured by the proportion of discriminations (i.e., the degree to a uniform distribution), reported that a normal distribution would be expected to have a discrimination of Delta >0.90.

Aims
To verify whether Delta is >0.90 when a sample with a normal distribution fits a Rasch (1960) model.

Sample
Rasch simulation data were used when sample sizes were 10, 50, 100, 200, 500, and 1000, item lengths were 5, 10, 20, 40, and 60, and 4 kinds of categories from 2 to 5 were manipulated in the study.

Methods
Sample data from normal and uniform distributions were yielded using a Rasch Rating Scale model. Ferguson's dichotomous Delta and Hankins' polytomous Delta_g were respectively produced. Another Delta setting a fixed number to 5 bins (Delta_5) was also computed for comparison with the former two. We simulated 100 times for those 24,000 (2 distributions × 6 samples × 5 item lengths × 4 categories) possible combinations and calculated 95 % confidence intervals (CIs) for the three aforementioned Delta values to verify whether Delta is >0.90 when a sample comes from a normal or a uniform distribution when the data fit a Rasch model.

Results
We found that (1) when samples are uniformly distributed and respond to a 2-point Rasch model test, Ferguson's dichotomous Delta = 0.96 (95 % CI = 0.86, 0.99), and Delta_5 = 0.94 (95 % CI = 0.80, 0.99). When responding to a polytomous Rasch model test, Delta_g = 0.97 (95 % CI = 0.88, 0.99), and Delta_5 = 0.96 (95 %

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

CI = 0.89, 0.99). (2) When samples are normally distributed and respond to a 2-point scale, Delta = 0.91 (95 % CI = 0.82, 0.97), and Delta_5 = 0.89 (95 % CI = 0.79, 0.96). When responding to a polytomous test, Delta_g = 0.94 (95 % CI = 0.86, 0.98), and Delta_5 = 0.89 (95 % CI = 0.80, 0.97).

Conclusion

There is insufficient evidence to expect that a normal distribution or a uniform distribution has a Ferguson's Delta >0.90 when considering its 95 % CI. For simple and easy use in the education field, we suggest that Delta_5 be used to describe the degree of equality of students' abilities within a class or between classes in a school.

PROMS2014TW009

Rasch Analysis of Patient and Nurse Assessed Satisfaction Gaps in Hospital Nursing Services

Presenter
Yu-Hui Huang, Tsair-Wei Chien, Weir-Sen Lin
Chiali Chi-Mei Hospital, Taiwan, Chia Nan University of Pharmacy and Science
E-mail: smile@mail.chimei.org.tw

Background

Many satisfaction surveys on patient perspective are conducted each year in a hospital. However, few focus on selecting for improvement the widest satisfaction gaps assessed by patients and nurses, especially those identified by the nurses.

Aims

To investigate the satisfaction gaps in nursing services assessed by patients and nurses in 3 different types of hospitals.

Keywords Satisfaction gap • Graphical representation • Rasch analysis • Mahalanobis distance • Euclidean Distance

Sample

Secondary data were collected from databanks of nursing service records and surveys conducted by a hospital group in southern Taiwan. These data were used for a series of statistical analyses on both scales responded to by nurses and patients in 2013. Fifty-six nursing units (medical center: 25; regional hospital: 24; local hospital: 7) were assessed by 1763 patients and 1270 nurses.

Methods

Two same-item questionnaires with 4-point Likert-type scales from the perspectives of patients and nurses were developed by healthcare experts in the studied hospital group. Rasch (1960) analysis was done using Winsteps software to investigate whether the scaled measurement was unidimensional. Parallel analysis was also used to determine the number of factors. Mahalanobis and Euclidean distance methods were used to answer the research question of similarity or difference in satisfaction, and a unique graphical was used to present it. Satisfaction gaps were determined using a graphical representation with the nursing unit means and their 95 % confidence interval (CI) of person and item parameters estimated using Rasch analysis.

Results

(1) Both scales were unidimensional and fit the Rasch model's expectations; (2) there were no significant differences in service satisfaction in the three hospitals; (3) five (8.9 %) nurse units expressed satisfaction gaps; (4) nurses expressed

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

significantly higher average satisfaction gaps than did patients' perception on three (30 %) items.

Conclusion

The unique graphical representation with 95 % CI that can be simply and clearly displayed the satisfaction gaps is recommended to present the differences in satisfaction in hospitals and/or on items in future.

An Aesthetic Intention Scale Using Rasch Analysis to Predict Whether Hospital Employees will Choose to Undergo Non-therapeutic Cosmetic Procedures

Presenter
Su-Chiu Fang, Tsair-Wei Chien
Chiali Chi-Mei Hospital, Taiwan
E-mail: smile@mail.chimei.org.tw

Background

To objectively and reproducibly assess the intention of aesthetic procedures remains one of the major, unmet challenges in plastic surgery. Frequently employed scoring systems for the evaluation of aesthetic intention and action are confounded by observer bias, be it that of the patient or of the surgeon. It is evident to us that Rasch model facilitates the objective, reproducible, standardized and specifically uniform evaluation of a scale by converting all ratings for any kind of aesthetic responses from a subjective value to an objective figure.

Aims

To develop a scale that predicts whether hospital employees will choose to undergo non-therapeutic cosmetic procedures and incorporates the so-called Rasch KIDMAP report card.

Keywords Plastic surgery • Aesthetic intention scale • KIDMAP • Unidimensionality • Parallel analysis

Sample

The setting was a 900-bed hospital in southern Taiwan. A total of 1,800 full-time workers in this hospital were asked to participate in a perception survey on aesthetic attitude in May 2009. The effective sample size was 1,124 for a return rate of 62.64 %.

Methods

Parallel analysis and exploratory factor analysis were used to determine the number of factors to retain. Rasch analysis with Winsteps software was used to verify the unidimensionality of the studied scale using the methods of principle component analysis of Rasch residuals and fit statistics. Multiple regression analysis was used to identify the key factors associated with aesthetic intention. A visual representation of the aesthetic intention so called Rasch KIDMAP was then prepared.

Results

The 12-item aesthetic intention scale has two characteristic features: subscales of perceived susceptibility and perceived need for cosmetic improvement, which are

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

fit to Rasch model's expectation, respectively. The two factors were evidence of elements that influence aesthetic intention. The report KIDMAP card can provide cosmetic surgeons and cosmetic surgery clinics with a tool for judging how likely one population subgroup, compared with others in a plot, is to choose non-therapeutic cosmetic procedures.

Conclusion

The aesthetic intention scale can be used to predict whether hospital employees will choose to undergo non-therapeutic cosmetic procedures.

PROMS2014TW014

The Booklet Design for a Longitudinal Study: Measuring Growth in New-Immigrant Children' Mathematics Achievement

Presenter
Pei-Jung Hsieh
Assistant Research Fellow, National Academy for Educational Research, Taiwan
E-mail: pjh@mail.naer.edu.tw

Due to the increased numbers of new-immigrant children in Taiwan, a large-scale assessment of mathematics achievement at the fourth grade was conducted in 2012 and a sample of panel members was followed in 2014. Hence, the study aimed to develop a valid instrument to measure and investigate the average new-immigrant students' growth between 4th and 6th grade in mathematical literacy. Total 78 multiple choice items into a set of 13 booklets are needed. The entire test items were selected from TASA 2012 (Taiwan Assessment of Student Achievement), which is the largest nationally representative and continuing assessment in Taiwan. Each item is designed to measure one of the four mathematics content areas, which are number and measurement, geometry, statistics and probability, and algebra. Three rules for identifying and selecting appropriate items are proposed. First, the number of anchor items is 39 (50 %), providing the basis for equating scores on different grades. Second, the 4th grade items with difficulty parameter larger than −0.5 and the 6th grade between −1 and 0.5 are considered as high priorities. Third, the proportions of each content area should be similar to the General Guidelines of Grades 1–9 Curriculum for Mathematics. Three-parameter logistic model was applied and PARSCALE was used to estimate item parameters. After these procedures, each test block contains six items and each booklet comprises four blocks by balanced incomplete block design. The mean difficulty of the items on each block ranged from 0.053 to 0.153. As a result, the mean difficulty of the items on each booklet ranged from 0.083 to 0.128. The items distributions across content areas are 64.10, 19.23, 5.13, and 11.54 %, mostly corresponding to the curriculum framework. This study demonstrated that detail consideration for the percent of anchor items, the range of item difficulties and the distribution of content areas could be useful for constructing measurement tool in a longitudinal study.

Keywords BIB design • Achievement growth • Panel study

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

PROMS2014UA001

The Implementation and Usage of Cloud Computing Among Pre-service Teacher: Validating Theory of Acceptance Model

Presenter
Dr. Enas Said Abulibdeh
Dean, Student Affairs
Assistant professor
Education Department, Al Ain University of Science and Technology
Al Ain, United Arab Emirates
E-mail: Enas220@gmail.com

The emerging technology cloud computing in Web 2.0 technology has attracted many researchers to investigate on its application in higher learning. However, there is still lacking of research in terms of application and the real usage in education context. This study seeks to investigate the adoption model of cloud computing using TAM by Davis (1989) with structural equation model analysis.

H1: ease of use significantly influence perceived usefulness
H2: perceived ease of use significantly influences behavioral intention to use
H3: perceived usefulness significantly influences behavioral intention to use
H4: behavioral intention to use significantly influences actual use

The study has been conducted for two semesters at Education College in AAU. The students were pre service teachers and Bachelor degree who were taking Educational Technology courses. They were exposed to the concept and technology of cloud computing. They were given 3 credit hours course which encompasses cloud computing usage from different types of technology of the Web 2.0 applications. At the end of the semester, students were given a set of questionnaire consisting of 30 self-constructed questions. This feedback enables the instructor cum researcher to identify the level of acceptance and to ensure the variables of usability in TAM predict the intention to use in their teaching practice in future. The total of students selected were purposive sampling of 239 registered for two semesters.

Thus, the student teachers have been exposed to the technology and applications prior to the research. Structural equation modeling (SEM) has been used for the analysis to estimate the model and paths. The findings have revealed the perceived ease of use impacted on the intention to use in future. Further, intention to use has been empirically proven in their actual use.

This study provides an implication to the educational context where cloud computing can be utilized in teaching the higher learning and school students. Further, the theory of acceptance model can be further researched in the context of

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

other emerging technologies. Specific technology requires specific acceptance variables in TAM. More variables need to be investigated further as it involves security, privacy and data integrity as forwarded by Truong (2010).

Keywords Structural equation model • Cloud computing • Perceived usefulness • TAM • Perceived ease of use

PROMS2014US001

Comparison of Rasch Scales Constructed from SF-36 Physical Functioning Items in Diverse Race Groups

Presenter
Joanne Wu
MD MS, Research Associate
Schaeffer Center for Health Policy and Economics, School of Pharmacy, University of Southern California
Los Angeles, USA
E-mail: qfw@usc.edu
Ningyan GU
Ph.D, Assistant Professor
College of Pharmacy, University of New Mexico
Albuquerque, USA
E-mail: ngu@salud.unm.edu

Background

SF-36 physical functioning scale (PF-10) is frequently used instruments for measuring self-reported physical function in clinical studies. However, the existing literature on racial differences in PF-10 is limited.

Aims

To compare item response patterns and Rasch scale in the PF-10 scale in diverse race groups.

Keywords Physical function • Rasch model • Item response patterns • Differential item function (DIF) • Race

Sample

The data set consisted of sociodemographic and survey data collected from a large managed care organization in the U.S. Participants (N = 9,967) who completed English version of the SF-36 and had no extreme responses in the PF-10 items were included in this study.

Method

PF-10 ask questions related to person's health limit in 10 activities, with 3 response choices (limited a lot, limited a little, and not limited at all). Rasch rating scale models were performed using Winsteps. Results were compared among different racial groups.

Results

The majority of the cohort was Caucasian (57.3 %) with a representative sample of African Americans (21.2 %), Latinos (13.3 %) and Asians (6.9 %). Person reliability ranged from 0.74 (Asian) to 0.83 (Caucasian and African American). Item

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

reliabilities were greater than 0.98. Two items (vigorous activities/bathing and dressing yourself) showed mis-fit (INFIT MNSQ were 1.20, 1.73 respectively; OUTFIT MNSQ were 3.50, 1.27 respectively) for all the ethnic groups. Items difficulty ranged from −3.04 logit for the item of bathing and dressing yourself to 3.73 logit for item of vigorous activities. All items showed DIF between two racial groups. However, the DIF had only a minimal impact on scale when all items were included in the scale measurement.

Conclusion

The study findings suggested that comparisons for physical function measured as a scale between racial groups are possible, although some items displayed DIF.

PROMS2014US006

Estimation of Non-compensatory Multidimensional Rasch Model

Presenter
Dr. Hong Jiao
Associate Professor
University of Maryland
Dr. Lihua YAO
Mathematical Statistician
Defense Manpower Data Center

Background

Due to estimation difficulty, non-compensatory multidimensional item response theory (MIRT) models are less frequently studied in research and applied in real practices. However, cognitive latent traits could be non-compensatory requiring the possession of multitraits simultaneously to answer an item correctly. For example, in a math test, if a student lacks of language proficiency to understand the question, though the student's math skill is very high, the probability of getting a correct response to the item could be extremely low.

Aims

The main objective of the study is to investigate the parameter estimation of the non-compensatory multidimensional Rasch model and the impact of misspecifying non-compensatory as compensatory models on ability parameter estimation.

Sample

This study is a simulation study. Sample sizes will be simulated at three levels: 500, 1000, and 2000.

Method

This simulation study will generate item response data based on the non-compensatory multidimensional Rasch model. Multiple factors will be manipulated including test length, sample size, the number of dimensions, and the correlation between latent traits. Test length will be simulated at two levels: 20 and 40 items; sample sizes at three levels: 500, 1000, and 2000. Two levels are simulated for the number of dimensions: two and three dimensions; and the correlations are simulated at four levels: 0, 0.3, 0.6, and 0.9. Both compensatory and non-compensatory multidimensional Rasch models are fitted with each data set. By fully crossing the levels of the manipulated factors, ninety-six study conditions will result. For each study condition, twenty replications will be implemented. Ability parameters are simulated from a multivariate normal distribution with means and variances are specified as standard normal distributions with respective correlations. Item difficulty parameters are each simulated from standard normal distributions as

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

well. Model parameter accuracy in terms of bias, standard error, and root mean square error will be evaluated and compared across models. BMIRT (Yao 2003) software program will be used for model parameter estimation for both compensatory and non-compensatory multidimensional Rasch models.

Results

This study is in progress. It is expected that the results will be ready by the mid of July.

Conclusions

The results and findings will be reported based on the full-scale simulation. The significance of the study lies in twofold. First, this study demonstrates the accuracy of model parameter recovery for the non-compensatory multidimensional Rasch model using BMIRT software. Second, the impact of misspecifying a non-compensatory model as compensatory will be demonstrated. The study results will inform the field that in selecting a multidimensional Rasch model, the compensatory model is not always the only choice, the non-compensatory Rasch model should be considered and the model parameters could be estimated using BMIRT.

PROMS2014US039

A Comparison of Two Measurement Approaches for Evaluating Standard Setting Ratings Within the Context of the AP World History Examination

Presenter
Yuk Fai Cheong
Associate Professor
Division of Educational Studies, Emory University, USA
Stefanie Wind
Emory University, USA
George Engelhard, Jr.
Professor
Department of Educational Psychology and Instructional Technology, University of Georgia, USA
Pamela Kaliski
Associate Psychometrician
The College Board, USA

Background

This study applied two different measurement approaches to examine the psychometric quality of the standard settings ratings assigned for the June 2012 Advanced Placement History (AP-WH) examination.

Aims

The analyses were guided by three major research questions:

(1) What are the major assumptions of the hierarchical cross-classified (HCM) (Rasbash and Goldstein 1994; Raudenbush 1993; Raudenbush and Bryk 2002) and the Rasch models (Engelhard 2009; Engelhard and Cramer 1997; Engelhard and Stone 1998; Rasch 1960/1980)?
(2) What do the two analyses reveal about the quality of the standard setting ratings assigned to the AP-WH items?
(3) What are the implications of the two sets of results for validity?

Keywords Cross-classified random effects models • Rasch Measurement Theory • Standard setting • Advanced Placement Examinations

Sample

Standard setting data used in this study were collected from the first two rounds (RND) of the standard setting procedures for the Advanced Placement World History (AP-HW) (The College Board 2011) in June, 2012. The ratings were assigned by fifteen panelists. Six of the panelists were high school AP World History teachers (HS), and nine were college-level world history instructors (Coll).

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*,
DOI 10.1007/978-3-662-47490-7

They had different prior experiences in scoring as well as years of teaching experience. During the two rounds, the panelists provided a rating for the cut scores on individual multiple-choice (MC) and constructed response (CR) items on the AP-WH exam.

Method

Analyses using the hierarchical cross-classified modeling (HCM) approach were conducted using the two-way cross-classified random effects models (Rasbash and Goldstein 1994; Raudenbush 1993). The Rasch measurement analysis was conducted using Facets (Linacre 2010).

Results

The HCM analyses showed that there was significant between-panelist and between-item variability for the standard setting ratings. Significant item type (MC vs. CR), significant round (RND) and level of experience (HS vs College) were detected. The Rasch analyses provided calibrations of the various facets (panelists, item, round, cut score, item and panelist characteristics) using the logit scale and revealed significant associations between panelist judgment with item type (MC vs. CR) and panelist subgroups (HS vs. Coll), as well as prior experience scoring AP exams, and years of teaching experience.

Conclusions

The two modeling approaches differ in their assumptions regarding how the panelist and the item effects are treated. In the HCM application, the models are specified to conceive of the effects associated with panelists and items as random. These effects are treated as fixed in the Rasch analyses. The results of the predictive effects of item type and panelist group are similar. The two approaches can both offer informative indices for evaluating the quality of the standard setting judgments, and the development of the AP-WH examination could likely benefit from the use of multiple types of quality assessment evidence.