Quan Zhang
Hong Yang *Editors*

Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding

Springer

Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding

Quan Zhang • Hong Yang
Editors

Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding

Springer

Editors
Quan Zhang
Hong Yang
Faculty of Foreign Studies
University of Jiaxing
Jiaxing, Zhejiang
China, People's Republic

ISBN 978-3-642-37591-0 ISBN 978-3-642-37592-7 (eBook)
DOI 10.1007/978-3-642-37592-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013940421

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

A Welcome Message

from
Chairman of PROMS
Professor Rob Cavanagh (PhD)
Chairman of the Board of Management
Pacific Rim Objective Measurement Society

On behalf of the Board of Management of the Pacific Rim Objective Measurement Society, it is my privilege and pleasure to welcome delegates to the 2012 Symposium in Jiaxing, Zhejiang Province, P.R. China. This is an auspicious occasion and marks the first time PROMS has been held on the Chinese mainland. I particularly need to sincerely thank Professor 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 in making PROMS happen. Also the members of the local committee, especially Professor Hong Yang and Professor Tao Danyu. The venue and organisation are both superb, we look forward to a highly enjoyable and intellectually stimulating program of social and academic events.

The notion of objective measurement, a core construct in Science and a guiding principle in the formation and operations of PROMS, concerns properties of measures. It is not necessarily the positivist view of nature and the world in which there exits an independent reality. Rather, it is about invariance, measures functioning independently of the location and time of their administration. For example, in the Human Sciences, a test of reading that assigns the same score to boys and girls with the same reading ability, to children from different countries with the same reading ability, and to children with the same reading ability tested at different times. A second feature is measurement being grounded in well-established theory, and data being collected to indicate attributes of persons or phenomena anticipated by this theory. This is in stark contrast to approaches that develop models from existing data, model-to-data fit. A third feature of objective measurement is the attention given to data on persons and attributes of the phenomenon of interest that do not

comply with the requirements of the measurement model. It is these anomalies that provide the impetus and focus of future research. When data fit a Rasch model, it is highly likely that these three features of measurement are manifest in the design and administration of a measure.

The foundation of PROMS and the annual Symposium are intended to inform researchers in the Pacific region about the practices and benefits of objective, and particularly, Rasch measurement. The Society supports applications in the fields of Business, Counselling, Economics, Education, Health Care, Language, Measurement, Psychology, Quality Assurance, Statistics and Strategic Planning. The Society advocates measurement practice that contributes to individual, community and societal well-being, and the public good, in the Pacific region.

The Society is supported by senior academics, researchers and scientists from Asia, Australia, North America and Europe. All share a vision of sustainable development in the Pacific that is informed by meaningful data and meaningful interpretations of these data. Also, they have a common concern about the inadequacies inherent in many decades of Western Human Science measurement and the potential for these shortcomings to be replicated elsewhere.

Welcome Message from the President of Jiaxing University

Xianmin Xu

Good morning!

Dear and respected Chairman ROB Cavanagh,

Dear and respected Professor A. Jackson Stenner, Professor Trevor Bond, Professor Magdalena Mo Ching Mok, Distinguished guests, ladies and gentlemen.

First of all, on behalf of Jiaxing University and on my personal behalf, I'd like to extend our warmest welcome to all the participants and VIP guests here today!

PROMS is the world's largest international objective measurement symposium mainly concerning Rasch Model. Over the past 8 years, it has been held in many parts of the Pacific Rim. Every year hundreds of experts and scholars attended the symposium. This year it has made its debut in Jiaxing, China Mainland. This shows that the domestic scholars and their foreign counterparts in the Rasch model field could have more opportunities to cooperate and communicate. It also provides an excellent platform of exchange and discussion for the application of Rasch Model and other related studies. I strongly believe the symposium will play a significant role in promoting China's research in Rasch model and other related fields.

Ladies and gentlemen, I'd like to take this opportunity to let you know something about our university. Jiaxing University has a long history, but it is the accredited institute of higher learning operating under Zhejiang Provincial Department of Higher Education ever since 2000. We adhere to the scientific development of our overall school strategic planning, which is characterized as "high-level, multi-layer, local-based, and application-oriented".

With the university motto "Act with integrity and probity, pursue academics industriously and rigorously" as its core, we are sparing no efforts to carry forward the "Red Boat Spirit" and creating a campus culture characterized with "Shiing-shen". Our graduate students are appreciated by the society and many alumni have become experts and celebrities in all walks of life across the country, contributing to the university's fame and far-reaching influence in the society.

Jiaxing, where our university is located, is a famous historical and cultural city of the country, the cradle of South China culture prevailing in the south of the Yangtze River, and it is also the birthplace of the Communist Party of China. With Shanghai in the east, Suzhou in the north, Hangzhou in the west and Hangzhou Bay in the south, Jiaxing is situated in the southeastern coast of China, which happens to be the center of the Yangtze Delta Area. Historically, Jiaxing was well developed in education and learning was greatly favored. The society was very peaceful with flourishing culture and prosperous economy. It had been recognized long before as "a land of abundance" as well as "the home of silk".

Professor Paul Ching-Wu Chu, a worldwide renowned physicist and the former President of Hong Kong Poly U, is now acting as the honorary president of Jiaxing University. Apart from this, we have one candidate for a Talent Project of New Century at national level, two specially invited professors who are titled as "Qianjiang Senior Talents" of Zhejiang Province. We also have one key innovation team, six distinguished teachers, four teaching teams all recognized at Zhejiang Provincial level.

Today we have more than 20,000 full-time undergraduate students, approximately another 10,000 registered adult students, about 100 long-term and short-term overseas students, and over 80 post-graduates under the joint master programs with other universities in Zhejiang Province and other provinces. Currently, we have more than 1,500 faculty members, of whom more than 1,000 are full-time teachers and more than 430 have senior professional titles, and 112 are full professors. Over 200 teachers are Ph.D. holders.

Our university has established 17 teaching units offering 45 four-year programs which cover nine categories of disciplines including economics, law, education, literature, science, engineering, medicine, management and art. We have two special construction projects at national level, eight provincial-level key disciplines and ten provincial-level key majors. Since 2006, we have undertaken 260 national-, provincial-, and ministry-level research projects, among which 16 projects have won provincial-, ministry-level rewards in teaching and scientific research. More than 1,600 papers have been published in core journals. More than 1,200 papers have been cited by the search systems such as SCI, EI, ISTP, etc.

Our university has three campuses; the total area is about 113.33 ha besides other 153.55 ha under planning. The total floor space of the university amounts to 620,000 square meters. Our picturesque campuses won the reputation of "Safe and Peaceful Campus" of Zhejiang Province, "Campus of Civilization" of Zhejiang Province, and "Advanced Unit in Landscaping" of Zhejiang Province. Apart from this, we have also established a modern experiment system with nine experimental centers as its basic framework, one experimental teaching demonstration center at national level, five provincial-level experimental teaching demonstration centers

and five municipal-level key laboratories. Our libraries hold over 1.48 million paper volumes as well as 1.2 million electronic volumes of books.

We encourage active participation in international academic exchanges and cooperation of various kinds so as to provide a wider platform for further development. So far we have established institutional collaborations with 23 universities in the USA, Germany, France, the UK, Denmark, Russia, South Korea and Japan, etc. We have joint master programs with foreign universities. We send students abroad for master degree and also for exchange students with partner universities. Each year we financially sponsor our faculties for visiting scholars abroad.

I was told that apart from our university, we have other sponsors who support this symposium in one way or other. They are MetaMetrics, Australia EAA, ETS, Springer, Beijing Foreign Language Teaching and Research Press, Higher Education Press and Shanghai Foreign Language Teaching Press. I'd like to express our heartfelt thanks to them all!

Finally, I wish PROMS2012 Jiaxing, China a complete success and I wish you all every success in your career. Have a nice stay in Jiaxing.

Thank you!

Welcome Message from Quan Zhang

from
Prof. Quan Zhang Ph.D.
Dean, College of Foreign Studies
Director, Institute of Language Testing
University of Jiaxing, Zhejiang Province, P.R. China

Entrusted by the Board of Management of the Pacific Rim Objective Measurement Symposium (PROMS), I'm very much delighted to announce that PROMS2012 will be held on campus of Jiaxing University, Zhejiang Province, P.R. China from August 6–9, 2012, with pre-conference workshops scheduled on 4–5 August, 2012, and post-conference self-arranged events scheduled on 10 August, 2012.

Over the past years, PROMS has been hosted in many parts of the Pacific Rim, in Singapore, Malaysia, Hong Kong, Taiwan and Tokyo, which has greatly promoted the research of and contributed to the development of Rasch Model in one way or another. It is the first time to be held in Jiaxing, China Mainland. Therefore, College of Foreign Studies, Jiaxing University is proud of that and are confident to offer good arrangement to make PROMS2012 more successful.

As early as in 1980s, the ideas and concepts regarding IRT was first introduced into China by Prof. Gui Shichun, my Ph.D. supervisor, and it is Prof. Gui who first conducted with great success the 10-year long (1990–1999) Equating Project for Matriculation English Test (MET) in China. MET is the most influential entrance examination for higher education administered annually to over 3.3 million candidates then. The Equating Project won recognition by Charles Alderson and other foreign counterparts during 1990s. Academically, those were Good Old Days

for Chinese testing experts and psychometricians. Then for certain reasons, the equating practice abruptly discontinued. Therefore, in China nowadays, the application of the IRT-based software like BILOG, Parscale, Winsteps, Iteman 4 and others to real testing problem solving is confined within an extremely small 'band' of people.

In this sense, I should say PROMS2012 thus meets an important need in that it provides an excellent introduction of IRT and its application. And anyone who is seriously interested in research and development in the field of psychometrics or language testing will find such a symposium and related workshops to be an excellent source of information about the application of Rasch Model.

PROMS2012 will follow the practice of the last PROMS, focusing on recent advances in objective measurement and providing an international forum on both the latest research in using Rasch measurement and non-Rasch practice. I'm sure the pre-conference workshops, parallel sessions, distinguished researchers and practitioners and post-conference self-arranged events will greatly encourage participation and all our participants would definitely share the benefits from it. In particular, I should mention that PROMS 2012 Jiaxing will benefit postgraduate students from developing countries and researchers who seek to use the Rasch Measurement in their research activities.

By the way, Jiaxing is a beautiful city in the Southern part of China Mainland. Here the lake, the bridge, the pavilion, the pagoda, the food, the people are inviting!

Acknowledgements

The following individuals and sponsors helped make the Pacific-Rim Objective Measurement Symposium (PROMS), held in Jiaxing, China, 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.

Robert, CAVANAGH	(AUS)
Trevor, BOND	(AUS)
Tetsuo, KIMURA	(Japan)
Magdalena Mo Ching, MOK	(Hong Kong)
A. Jackson, STENNER	(USA)
William P., Jr., FISHER	(USA)
Wenchung, WANG	(Taiwan)
Eric, WU	(USA)
Quan, ZHANG	(China)
Hong, YANG	(China)
Danyu, TAO	(China)
Yulei, QIU	(China)
Chun-yan, ZHU	(China)
Ming-zhu, MIAO	(China)
Huifen, MU	(China)
Lan, MO	(U.K.)
Yan, ZHAO	(China)
Xiaolian, FU	(China)
Guo-Xiong, HE	(China)
Lan-lan, DING	(China)
Feng, CHEN	(China)
Yan-fei, WU	(China)
Zi, YAN	(Hong Kong)

Sponsors:

University of Jiaxing, Zhejiang Province, China
Higher Education Press, Beijing, China
Foreign Language Teaching and Research Press, Beijing, China
Springer, Germany
MetaMetrics, USA
Educational Assessment Australia, Australia
Educational Testing Service, USA
Taylor & Francis, UK

Conference Organizer

Conference Chair
Prof. ZHANG Quan
Professor, Faculty of Foreign Studies, University of Jiaxing

Organizing Committee

Co-Chairs:
Prof. ZHANG Quan
Professor, Faculty of Foreign Studies, University of Jiaxing

Prof. YANG Hong
Professor, Faculty of Foreign Studies, University of Jiaxing

Prof. TAO Danyu
Professor, Faculty of Foreign Studies, University of Jiaxing

Members:
MU Huifeng
Faculty of Foreign Studies, University of Jiaxing

QIU Yulei
Faculty of Foreign Studies, University of Jiaxing

HE Guoxiong
Faculty of Foreign Studies, University of Jiaxing

MIAO Mingzhu
Faculty of Foreign Studies, University of Jiaxing

ZHU Chunyan
Faculty of Foreign Studies, University of Jiaxing

MO Lan
Executive Officer, Faculty of Foreign Studies, University of Jiaxing

ZHAO Yan
Faculty of Foreign Studies, University of Jiaxing

FU Xiaolian
Faculty of Foreign Studies, University of Jiaxing

DING Lanlan
Faculty of Foreign Studies, University of Jiaxing

CHEN Feng
Faculty of Foreign Studies, University of Jiaxing

WU Yan-fei
Faculty of Foreign Studies, University of Jiaxing

PROMS 2012 Jiaxing,China
August 6-9, 2012

Sponsors

Contents

Chapter 1
On the Potential for Improved Measurement in the Human and Social Sciences

William P. Fisher Jr. and A. Jackson Stenner

Abstract Geometry is the most ancient branch of physics. All linear measurement is essentially a form of practical geometry. Following Maxwell's method of drawing analogies from geometry, Rasch conceptualized measurement models as analogous to scientific laws. Rasch likely absorbed Maxwell's method via close and prolonged interactions with colleagues known for their use of it. Examination of the common form of the relationships posited in the Pythagorean theorem, multiplicative natural laws, and Rasch models leads to a new perspective on the potential unity of science. To be fully realized in the social sciences, Rasch's measurement ideas need to be dissociated from statistics and IRT, and instead rooted in the Maxwellian sources Rasch actually drew from. Following through on the method of analogy from geometry may make human and social measurement more intuitive and useful.

Keywords Geometry • Measurement • Scientific law • Rasch models

1.1 Introduction

All linear measurement makes use of the geometric figure of the line. For persons educated in basic scientific conventions, quantitative comparisons automatically bring images of a number line to mind. Despite these associations, most statistical methods in the social sciences do not require experimental tests of the hypothesis that any given numeric difference stands for a constant unit amount. Further, to many the very idea

W.P. Fisher Jr. (✉)
Graduate School of Education, University of California, Berkeley, CA, USA

LivingCapitalMetrics.com, Sausalito, CA, USA
e-mail: wfisher@berkeley.edu

A.J. Stenner
MetaMetrics, Inc., Durman, NC, USA

University of North Carolina, Chapel Hill, NC, USA

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_1,

that geometry could provide a useful basis for measurement in the social sciences seems implausible. To what extent, however, might this implausibility be more a function of unexamined prejudices than careful reasoning? There may be more of value in this line of thinking than meets the eye.

1.2 Linear Measurement as Practical Geometry

In the natural sciences, the basis for quantitative units is established, in effect, via analogies from geometry. The Pythagoreans considered tonal proportions to be the geometry of motion, for instance, encompassing sound, celestial bodies, and the human soul in a comprehensive cosmology (Isacoff 2001, p. 38). Similarly, the essential question for Copernicus was not "Does the earth move?" but, rather, ". . .what motions should we attribute to the earth in order to obtain the simplest and most harmonious geometry of the heavens that will accord with the facts?" (Burtt 1954, p. 39). Both Boscovich and Legendre based their contributions to the method of least squares in geometrical formulations (Stigler 1986, pp. 42, 46, 47, 57). Galileo "derived his rule relating time and distance using geometry" (Heilbron 1998, p. 129). Einstein (1922) considered geometry to be "the most ancient branch of physics," according "special importance" to his view that "all linear measurement in physics is practical geometry," "because without it I should have been unable to formulate the theory of relativity" (p. 14).

Though the method of least squares is foundational to contemporary statistical analysis, it was originally formulated by Boscovich, who "followed in a Newtonian tradition of giving geometric descriptions rather than analytic ones" (Stigler 1986, pp. 42–43, 51). Boscovich's work was only later expressed analytically, by Laplace. Pledge (1939) makes the historical connection between geometry and natural law in the general point that

> as the Greeks gave us the abstract ideas (point, line, etc.) with which to think of space, and the 17th century those (mass, acceleration, etc.) with which to think of mechanics, so Carnot gave us those needed in thinking of heat engines. In each case the ideas are so pervasive that we use them even to state that they never apply exactly to visible objects (p. 144).

Narens (2002) explicitly roots measurement theory in a Pythagorean sense of scientific definability focused on meaningfulness as invariance across transformations. Maxwell provides the clearest method for making linear measurement analogous with practical geometry (Black 1962; Nersessian 2002; Turner 1955). Inventing the contemporary concept of mathematical modeling (Hesse 1961, p. 206), Maxwell freed physics from the constraints of Newtonian mechanics via his concept of the abstract mathematical field (Rautio 2005, p. 53; McMullin 2002). His work still stands as one of the most productive examples of how to draw geometric analogies of phenomena (Klein 1974, p. 474; Rautio 2005).

To understand Maxwell's method of analogy, it is important to know that, in the eighteenth and nineteenth centuries, scientists and philosophers in many fields employed Newton's laws of motion as a framework for structuring investigations

of a wide range of different phenomena. Newton's theory of gravitation provided the form of a Standard Model adopted across the sciences of nature as the hallmark criterion of scientific method (Heilbron 1993, pp. 5–6).

Nersessian (2002) concurs, saying "After Newton, the inverse-square-law model of gravitational force served as a generic model of action-at-a-distance forces for those who tried to bring all forces into the scope of Newtonian mechanics" (p. 139). Maxwell learned the method of drawing analogies from the standard model from his colleague William Thomson (Lord Kelvin), and told him that he "intended to borrow it for a season. . .but applying it in a somewhat different way" (Nersessian 2002, p. 144).

The difference between Thomson's method and Maxwell's use of it is telling. Like Maxwell, Thomson constructed a number of analogies, such as between heat and electrostatics. But Thomson merely took existing equations describing a known physical system and changed the names of the parameters to match the system under investigation (Nersessian 2002, p. 144). This was the typical way in which the Standard Model was applied in research up to that time.

The superficiality of this method, however, made it vulnerable to two errors Maxwell (1965/1890, p. 155) sought to avoid, distraction by abstract mathematical analyses and by too-literal preconceptions of the physical phenomenon. As Maxwell put it,

> By referring everything to the purely geometrical idea of the motion of an imaginary fluid, I hope to attain generality and precision, and to avoid the dangers arising from a premature theory professing to explain the cause of the phenomena. . . [so that one might in due course arrive at] a mature theory, in which physical facts will be physically explained (Maxwell 1965/1890, p. 159).

Maxwell (1965/1890, p. 155) considered a too-quick leap to mathematical analysis a distraction, saying purely mathematical simplifications are likely to cause the investigator "entirely lose sight of the phenomena to be explained; and though we may trace out the consequences of given laws, we can never obtain more extended views of the connexions of the subject." In the human and social sciences, little attention is paid to modeling constructs, though there are several significant exceptions (Burdick et al. 2010; Dawson et al. 2006; Stenner et al. 1983; Wilson 2005, 2008) that take up the challenge in ways analogous to the approach advocated by Maxwell, in terms of psychosocial explanations of psychosocial facts.

Maxwell, then, started from simple geometric ideas and built up an understanding of the construct via analogy (Black 1962; Nersessian 2002; Turner 1955). In so doing, he provided "the prototype for all the great triumphs of twentieth-century physics" (Dyson, in Rautio 2005, p. 53). Ludwig Boltzmann considered Maxwell's method of analogy as important as his scientific work (Boumans 2005, pp. 24, 28). Boltzmann's student, Ehrenfest, and Ehrenfest's student, Tinbergen, each employed Maxwell's approach to mathematical modeling and his method of analogy in their studies in economics (Boumans 2005, pp. 24, 28, 31, 41).

Rasch was, then, connected through his associations with Tinbergen, Frisch, and Koopmans (Frisch's and Tinbergen's student) with a direct line of intellectual descent from Maxwell (Fisher 2010). Rasch (1960, pp. 110–115) established a

basis for a Maxwellian Standard Model in the social sciences when he structured his models in the pattern of Maxwell's analysis of mass, force, and acceleration. Few researchers to date, however, have noted or expanded upon the connection Rasch drew between his models and Maxwell's analysis, in large part because Rasch himself did not effectively follow through to a full implementation of Maxwell's method. The quality of research using Rasch's models suffers for this loss.

Rasch presented his models in a manner similar to Thomson's method of merely substituting parameter names across the different phenomena studied, and this is, in effect, exactly how Rasch models are usually applied. Easily performed computer analyses disconnect statistical considerations from the conceptualization and evaluation of the construct (Stenner et al. 1983; Wilson 2013). The question then arises as to how a shift from Thomson's method to Maxwell's might be achieved in the human and social sciences.

Significant untapped potential for such a shift can be found in the shared mathematical formalism of the Pythagorean theorem, the multiplicative structure of natural laws, and Rasch models. These connections suggest much could be gained from closer study of Maxwell's reasoning process (Nersessian 2002) and the ways in which it is similar to and different from predictive construct models.

1.3 Geometry and Natural Law

Figure 1.1 illustrates a proof of the Pythagorean theorem, where the square of the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides:

$$a^2 + b^2 = c^2$$

For Fig. 1.1, this works out as:

$$3^2 + 4^2 = 5^2 = 9 + 16 = 25$$

Most scientific laws are, however, written in a multiplicative form (which also includes equations involving division) (Crease 2004; Taagepera 2008; Burdick et al. 2006), like this:

$$a = f/m$$

or

$$f = m * a$$

where the acceleration of an object can be estimated by dividing the applied force by the object's mass, or the force is estimated by multiplying the mass by the

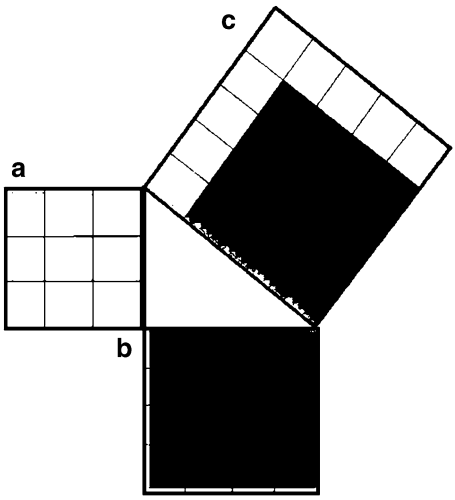


Fig. 1.1 A proof of the Pythagorean theorem

acceleration. This, of course, is how Maxwell (1920/1876) presented Newton's Second Law.

Other geometric relationships have the same multiplicative form as scientific laws, such as the definition of the circle as a closed arc equidistant from a single point, with the circumference equal to pi times the radius squared. The Pythagorean theorem can also be written in the form of a multiplicative law, by means of the number e (2.71828...) (Maor 1994):

$$e^9 * e^{16} = e^{25}$$

Substituting a for e^9, b for e^{16}, and c for e^{25} in this description of the triangle in Fig. 1.1 gives:

$$\mathrm{a} * \mathrm{b} = \mathrm{c}$$

and could be solved as

$$8103 * 8,886,015 \approx 72,003,378,611$$

Converting back to the additive form using the natural logarithm, the equation looks like this:

$$\ln(8,103) = \ln(72,003,378,611) - \ln(8,886,015)$$

and this

$$9 = 25 - 16.$$

Whether expressed in multiplicative or additive forms, Newton's Second Law and the Pythagorean theorem both define the way changes in one parameter in a mathematical model result in proportionate changes in the other parameters.

Furthermore, the empirical relational structure stays the same no matter what unit characterizes the numerical relational structure. Maxwell presented Newton's Second Law in this form:

$$A_{vj} = F_j / M_v.$$

Applying catapult j's force F of 7.389 N (53.445 poundals) to object v's mass M of 1.6487 kg (3.635 lb) results in an acceleration of 4.4817 m (14.70 ft) per second, per second. (That is, 7.389/1.6487 = 4.4817, or 53.445/3.635 ≈ 14.70). The proportional relationships are constant no matter which units are used, satisfying the criterion of meaningfulness (Mundy 1986; Narens 2002; Rasch, 1961). In this context, Rasch (1960, 112–113; Burdick et al. 2006) noted that,

> If for any two objects we find a certain ratio of their accelerations produced by one instrument, then the same ratio will be found for any other of the instruments. Or, in a slightly mathematized form: The accelerations are proportional.

Conversely, it is true that if for any two instruments we find a certain ratio of the accelerations produced for one object, then the same ratio will be found for any other objects.

Rasch's (1961, p. 322) model for measuring reading ability and text reading difficulty has the multiplicative form of

$$\varepsilon_{vi} = \theta_v \sigma_i$$

and the additive form (Rasch 1961, p. 333):

$$\varepsilon_{vi} = \theta_{v+} \sigma_i.$$

Rasch (1960, pp. 110–115) cites Maxwell's presentation of Newton's Second Law as his source for these formulations. This model takes reading comprehension ε as the product (or the sum) of person v's reading ability θ and item i's text complexity σ. The model is also often written as

$$\Pr\{X_{ni} = 1\} = e^{\beta n - \delta i} / 1 + e^{\beta n - \delta i}$$

or

$$P_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$$

or

$$\ln[P_{ni} / (1 - P_{ni})] = B_n - D_i$$

which is to say that the log-odds of a correct response from person n on item i is equal to the difference between the estimate B of person n's ability and the estimate D of item i's difficulty (Wright 1997; Wright and Stone 1979). Moving the effect of e from one side of the equation to the other makes the response odds equal to e taken to the power of the difference between B and D, divided by one plus e to that power.

In light of the proportionality obtained in these relationships, Rasch (Rasch 1960; also see his 1961, p. 325) formulated a separability theorem in terms that apply to both additive and multiplicative forms of the models, saying

> It is possible to arrange the observational situation in such a way that from the responses of a number of persons to the set of tests or items in question we may derive two sets of quantities, the distributions of which depend only on the test or item parameters, and only on the personal parameters, respectively. Furthermore, the conditional distribution of the whole set of data for given values of the two sets of quantities does not depend on any of the parameters (p. 122).

The separability of the parameters is evident in the proportionality of the relationships expected by the model. As any one parameter is varied relative to a second parameter, values for the third are predictable. For example, for a person-item interaction in which there is a 0.82 likelihood of a correct response, the odds ratio of 4.556 (0.82/0.18) gives a log-odds (logit) difference of 1.5 between the person ability and item difficulty estimates (see Wright and Stone 1979, p. 16, for a table relating response probabilities to logit differences). Any ability measure that is 1.5 logits different from a difficulty calibration implies a 0.82 probability of a correct response.

If the 1.5 logit difference results from a comparison of a person measure of 2.0 and an item calibration of 0.5, then, to obtain the multiplicative form of the model,

$$\varepsilon_{vi} = \theta_v \sigma_i$$

we have, with the previous values entered

$$e^{2.0} = e^{1.5} * e^{0.5},$$

which is exactly the same equation as that previously used to illustrate Newton's Second Law: $7.389 = 4.4817 * 1.6487$.

1.4 Predictive Construct Modeling

Rasch (1960, 2010/1972) explained how the structure of Newton's second law of motion (relating force, mass, and acceleration) is analogous to the structure of a law relating reading ability, text complexity, and comprehension rates. Rasch held that,

> Where this law can be applied it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus. . .the reading accuracy of a child . . . can be measured with the same kind of objectivity as we may tell its weight. . .. (p. 115)

Wright (1997, p. 44), a physicist who worked with Nobelists Townes and Mulliken before turning to psychology and collaborations with Rasch, concurs, saying, "Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics." Andrich (1988, p. 22) observes that "...when the key features of a statistical model relevant to the analysis of social science data are the same as those of the laws of physics, then those features are difficult to ignore."

In his retirement speech, after describing multiple examples and elaborating the logic of the analogy in detail, as he also had in his book (Rasch 1960, pp. 110–115), Rasch (2010/1972) concluded that,

> With all of this available to us, we will have an instrumentarium with which many kinds of problems in the social sciences can be formulated and handled with the same types of mathematical tools that physics has at its disposal—without it becoming a case of superficial analogies (p. 1272).

But nowhere in his book, retirement lecture, or other publications does Rasch provide a theory of a substantive construct behaving in accord with the structure of a lawful regularity. As Maxwell understood would happen, the convenient analytical formulation of Rasch's models has caused us to lose sight of the phenomena to be explained, such that we "never obtain more extended views of the connexions of the subject" (Maxwell 1965/1890, p. 155). Rasch emphasized the positing and testing of invariances, but ignored the constitutive cause and effect relationships.

In asserting that "Thereby you can gradually reach a clarification of the field of validity of the law," and in next taking "a closer look at the contents of the law," Rasch (2010/1972, p. 1254) does not follow Maxwell's process. Rasch does not try to explain individual-centered variation in a psychological or social phenomenon in psychological or social terms, as one would in investigations emulating Maxwell's interest in explaining a physical phenomenon in physical terms. Instead, Rasch's focus on the contents of the law is strictly mathematical. His concern is with the nature of the independence of the comparisons made in a context of infinite possibility. He shows how the frame of reference provides a means for defining all possible relevant observational situations, but he does not show, as does Maxwell for electromagnetism, what makes any given observation conform to the model in the way that it does.

In the wake of Rasch's work and later large-scale studies equating high stakes reading tests (Jaeger 1973; Rentz and Bashaw 1977), however, Stenner and colleagues (Stenner 2001; Stenner et al. 2006) developed an effective and parsimonious predictive theory of what makes text easy or difficult to read. Others have similarly devised predictive models of other cognitive and behavioral constructs (Dawson et al. 2006; Embretson 1998; Fischer 1973; Fisher 2008; Green and Kluever 1992; Wilson 2008) with the aim of achieving the degree of control over the instrumentation needed for the reliable and highly efficient automated production of assessment items (Bejar et al. 2003; Stenner and Stone 2003).

Generalizing these accomplishments requires a systematic and methodical way of interweaving substantive qualitative content and abstract mathematical construct

issues. Various systems for assessing constructs (Embretson 1998; Stenner and Smith 1982; Stenner et al. 1983; Burdick et al. 2010; Wilson 2005) set the stage for fuller realizations of model-based reasoning in the psychosocial sciences by prioritizing theory development. In the context of these systems, hypotheses are formulated and tested by iterating through a sequence of moments in a method, any one of which may serve as a point of entry or exit. Building on the way in which data, instruments, and theory have each historically served to mediate each other's interrelations in the history of science (Ackermann 1985), and focusing on the predictive control of the construct, new horizons for qualitatively-informed quantitative social science can be envisioned.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton: Princeton University Press.

Andrich, D. (1988). *Rasch models for measurement* (Sage University paper series on quantitative applications in the social sciences, Vol. 07–068). Beverly Hills: Sage Publications.

Bejar, I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2*(3), 1–29.

Black, M. (1962). *Models and metaphors*. Ithaca: Cornell University Press.

Boumans, M. (2005). *How economists model the world into numbers*. New York: Routledge.

Burdick, D. S., Stone, M. H., & Stenner, A. J. (2006). The combined Gas Law and a Rasch Reading Law. *Rasch Measurement Transactions, 20*(2), 1059–1060.

Burdick, D. S., Stenner, A. J., & Kyngdon, A. (2010, Summer). From model to measurement with dichotomous items. *Journal of Applied Measurement, 11*(2), 112–121.

Burtt, E. A. (1954). *The metaphysical foundations of modern physical science* (Rev. ed.). Garden City: Doubleday Anchor.

Crease, R. (2004). The greatest equations ever. *Physics World, 17*(10), 19.

Dawson, T. L., Fischer, K. W., & Stein, Z. (2006). Reconsidering qualitative and quantitative research approaches: A cognitive developmental perspective. *New Ideas in Psychology, 24*, 229–239.

Einstein, A. (1922). Geometry and experience. In G. B. Jeffery, W. Perrett (Trans.), *Sidelights on relativity* (pp. 12–23). London: Methuen & Co., Ltd.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fisher, W. P., Jr. (2008). *A predictive theory for the calibration of physical functioning patient survey items*. Presented at the second conference on patient reported outcome measurement information systems, Bethesda: NIH and NIAMS, March 2–5.

Fisher, W. P., Jr. (2010). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics: Conference Series*, *238*(1). http://iopscience.iop.org/1742-6596/238/1/012016/pdf/1742-6596_238_1_012016.pdf.

Green, K. E., & Kluever, R. C. (1992). Components of item difficulty of Raven's matrices. *The Journal of General Psychology, 119*, 189–199.

Heilbron, J. L. (1993). Weighing imponderables and other quantitative science around 1800. *Historical Studies in the Physical and Biological Sciences*, *24*(Supplement), Pt. I, 1–337.
Heilbron, J. L. (1998). *Geometry civilized: History, culture, and technique*. Oxford: Clarendon.
Hesse, M. (1961). *Forces and fields: A study of action at a distance in the history of physics*. London: Thomas Nelson and Sons.
Isacoff, S. M. (2001). *Temperament: The idea that solved music's greatest riddle*. New York: Knopf.
Jaeger, R. M. (1973). The national test equating study in reading (The Anchor Test Study). *Measurement in Education, 4*, 1–8.
Klein, H. A. (1974). *The world of measurements: Masterpieces, mysteries and muddles of metrology*. New York: Simon & Schuster.
Maor, E. (1994). *E: The story of a number*. Princeton: Princeton University Press.
Maxwell, J. C. (1920/1876). *Matter and motion* (J. Larmor, Ed.). New York: The Macmillan Co.
Maxwell, J. C. (1965/1890). *The scientific papers of James Clerk Maxwell* (W. D. Niven, Ed.). New York: Dover Publications.
McMullin, E. (2002). The origins of the field concept in physics. *Physics in Perspective, 4*(1), 13–39.
Mundy, B. (1986). On the general theory of meaningful representation. *Synthese, 67*(3), 391–437.
Narens, L. (2002). A meaningful justification for the representational theory of measurement. *Journal of Mathematical Psychology, 46*(6), 746–768.
Nersessian, N. J. (2002). Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.), *Essays in the history and philosophy of science and mathematics* (pp. 129–166). Lasalle: Open Court.
Pledge, H. T. (1939). *Science since 1500: A short history of mathematics, physics, chemistry, biology*. London: His Majesty's Stationery Office.
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: Danmarks Paedogogiske Institut.
Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Contributions to biology and problems of medicine, Vol. IV, pp. 321–333). Berkeley: University of California Press.
Rasch, G. (2010/1972). Retirement lecture of 9 March 1972: Objectivity in social sciences: A method problem (C. Kreiner, Trans.). *Rasch Measurement Transactions*, *24*(1), 1252–1272.
Rautio, J. C. (2005). Maxwell's legacy. *IEEE Microwave Magazine, 6*(2), 46–53.
Rentz, R. R., & Bashaw, W. L. (1977). The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement, 14*(2), 161–179.
Stenner, A. J. (2001). The Lexile Framework: A common metric for matching readers and texts. *California School Library Journal, 25*(1), 41–42.
Stenner, A. J., & Stone, M. (2003). Item specification vs. item banking. *Rasch Measurement Transactions, 17*(3), 929–930.
Stenner, A. J., & Smith, M., III. (1982). Testing construct theories. *Perceptual and Motor Skills, 55*, 415–426.
Stenner, A. J., Smith, M., III, & Burdick, D. S. (1983, Winter). Toward a theory of construct definition. *Journal of Educational Measurement, 20*(4), 305–316.
Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement, 7*(3), 307–322.
Stigler, S. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge: Harvard University Press.
Taagepera, R. (2008). *Making social sciences more scientific: The need for predictive models*. New York: Oxford University Press.
Turner, J. (1955). Maxwell on the method of physical analogy. *The British Journal for the Philosophy of Science, 6*, 226–238.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah: Lawrence Erlbaum Associates.
Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift Für Psychologie/ Journal of Psychology, 216*(2), 74–88.
Wilson, M. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika, 78*(2), 211–236.
Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45, 52.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Chapter 2
A Pilot Study Based on Rasch into the Appropriateness of the TOEIC Bridge Test for Chinese Students: Status Quo and Prospect

Quan Zhang, Mingzhu Miao, Chunyan Zhu, and Eng Han Tan

Abstract The present research reports carefully the situation of language testing for vocational students in China. As China is now open wider to the outside world, vocational students, after graduation, are required to use more and more English for international communication. Such an ability to communicate in English, the listening comprehension in particular, has become essential for success in business, workplace and life across China as well as around the world. Therefore, the authors point out that China needs a more objective measure of English proficiency for lower-intermediate-level learners particularly the vocational students. Based on this, the appropriateness of TOEIC Bridge Test focusing on listening comprehension part to Chinese vocational students as a measure of English listening comprehension skill was evaluated using correlations between the scores of listening comprehension part of TOEIC Bridge Test and that of a local English test followed by a questionnaire to confirm the validity of the scores thus obtained. The purpose of conducting such a study is to explore the feasibility of using TOEIC Bridge test in place of some local tests. The research report was submitted to the concerned department of higher education for reference. Apart from this, some common practices under the principles of language testing theory like IRT are also touched upon for popularization.

Keywords TOEIC Bridge Test • Correlations • Rasch • GiTEST

This project is financially supported by the Supervisory Committee of ELT in Vocational Higher Education, Ministry of Education, P.R. China in 2010–2011.

Q. Zhang (✉) • M. Miao • C. Zhu
Faculty of Foreign Studies, University of Jiaxing, Jiaxing, China
e-mail: gzjohnzh@gmail.com

E.H. Tan
Educational Testing Service (Beijing Office), Princeton, NJ, USA

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_2,

2.1 Introduction

Today in China, the number of Chinese vocational students enrolled each year is increasing. Up to the present, the total number of 1,795 vocational colleges nationwide amounts to almost 3.3 millions, of whom approximately 1.5 millions are English majors. However, over the past decades, all the students were taking a national English test, a compulsory test administered twice a year to such a big number of students across China. As the test certificates have been found out not very well recognized by both society and educators in China, an official document issued by Ministry of Education, China in 1996 stipulates that the test scores of such a national test were no longer taken as the reference for evaluation purpose of the English teaching quality operated by vocational colleges across China.

Because of this, the number of test takers is decreasing annually. Many vocational colleges are seeking other ways out. The situation can be compared as "Three Kingdoms" of English Tests Criteria. The key issue, as far as tests are concerned, is not very well recognized. This of course raises the issues of the validity of the test. Regardless of this, the test is still administered. Therefore, the present situation is actually going like this: all of these 3.3 millions of vocational students, some still take the national test, others take TOEIC Bridge test instead, and still some others even take a test designed at their own provincial level.

2.2 TOEIC Bridge Test

Among many English language tests developed and administered by Educational Testing Service (ETS) today, TOEIC *Bridge*™ test (http://www.ets.org/toeicbridge) stands for the "Test of English for International Communication" and is thus paid attention to by both Chinese government and university educators for at least four reasons. TOEIC Bridge Test is officially introduced by Ministry of Labor and Social Security (MOLSS) under Chinese government. According to reliable sources (Ashmore et al., 2009), it is the choice of more than 180,000 examinees a year and is recognized by hundreds of corporations worldwide. The format of TOEIC *Bridge*™ is 1-h paper-and-pencil test of 100 multiple-choice questions divided into two parts: listening comprehension and reading comprehension. Hence, it is easy to administer. Apart from this, both TOEIC *Bridge*™ test and the Chinese test help verify everyday beginning and lower-intermediate English-proficiency levels for communication in English at the workplace. And finally, the test takers for TOEIC Bridge Test may be students of English or people whose native language is not English and who need to use English for work or travel, or students who are learning English and are at a beginning to lower-intermediate level learners of English who are taking commercial English language courses. All these coincide in formality with that of the local Chinese test.

2.3 The Research Purpose

To validate the test scores from the TOEIC Bridge as a measurement of English proficiency for and to eventually replace some local tests currently administered to Chinese vocational students, the *Supervisory Committee of EFT in VHEME*[1] decided to initialize the pilot validity study. In 2008, the first TOEIC Bridge Test was successfully administered by Educational Testing Service to 4,050 registered students of ten sampled (by the Committee) vocational institutes across China. And the second test was done in the same way the following year. The data were provided by ETS. Then some proposal for further improvement was soon put forward. Would it be possible to compare the scores obtained by students taking the two tests? Therefore, apart from the TOEIC Bridge test, the sampled subjects in Guangdong and Zhejiang Province were also given a kind of local test. The present study is part of this project focusing on listening comprehension. The purpose of conducting such a study is in an attempt with academic honesty to achieve the goal regarding the appropriateness of TOEIC Bridge Test to Chinese vocational students plus a research report to department of higher education for reference and decision-making. The authors also took the research by Siharay et al. (2009) as reference.

2.4 Research Design

The present research is a pilot study based on Rasch Model into the Appropriateness of the TOEIC Bridge Test for Chinese students. We use GiTEST, a kind of Rasch-based software to process all the data. GiTEST assumes binary (right-wrong) scoring. Designed for applications of both CTT and IRT theory to practical testing problems, it can be held as the earliest application of Rasch Model in China. From 1990 to 1999, GITEST was used to undertake 10-year Equating Project of Matriculation English Test (MET) launched by the Examination Authority under China Ministry of Education.

2.4.1 Subjects

Students of Jiaxing University and other universities within Zhejiang and other provinces in China are used as subjects. The data to be shown here are based on those of Jiaxing University and a university in Ningbo of Zhejiang Province. Totally, we had 110 subjects.

[1] Vocational Higher Education, Ministry of Education, P.R. China

2.4.2 *Method*

Designed to test the concurrent validity (See, e.g. Bachman and Palmer 1983, and concurrent and predicative validity by Hughes, 1989; Zhang, 2004; Zhang, 2011, et al.) and for a better comparison of TOEIC Bridge Test and the relevant English test practised in China, a test with mixed test items of both TOEIC Bridge Test and the Local test was administered to a group of vocational students sampled in Zhejiang Province. We don't follow the traditional practice to administer two separate tests to our subjects. Instead, we mixed test items of both tests and administer only one test so as to ensure that the data to be collected will be more reliable as students taking the test will give equal attention to all the test items they are coping with during the whole test performance. GiTEST is used to process all the item analyses, test scoring, correlation and comparison.

To further confirm the results, an immediate post-test questionaire was distributed to each of the test takers. The results thus obtained show that TOEIC Bridge test fits Chinese students better. Evidences collected from questionaires indicate a higher correlation between the TOEIC Bridge scores and the test takers' communicative competence in the real social and campus life.

2.5 Results and Analysis

In this section, we are addressing the problems demonstrated in two figures based on the data processed by GiTEST. Figure 2.1 shows the scores of listening comprehension of TOEIC Bridge and Chinese test takers' communicative competence in the real social and campus life, and Fig. 2.2, the scores of listening comprehension of TOEIC Bridge and that of listening comprehension of the local Chinese test. Let's examine Fig. 2.1 first.

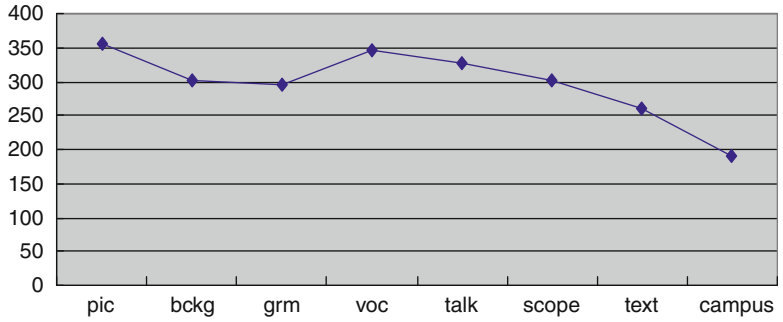


Fig. 2.1 The scores of listening comprehension of TOEIC Bridge and Chinese test takers' communicative competence in the real social and campus life. (N = 50) based on the data collected from questionaires

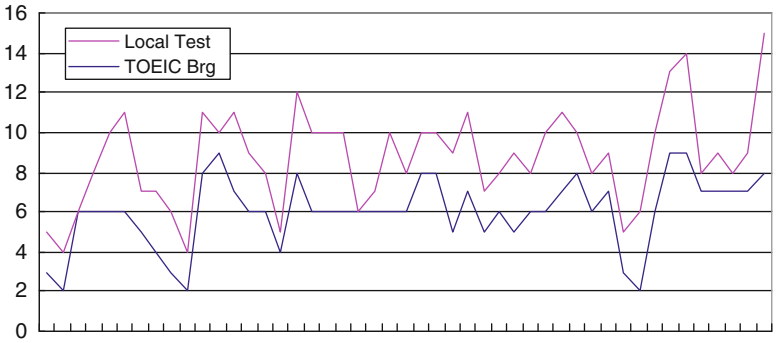


Fig. 2.2 The scores of listening comprehension of TOEIC Bridge and that of listening comprehension of the local Chinese test. (N = 110) Correlate coefficient = 0.146; Mean TOEIC Bridge = 6; Mean local test = 2.77; Mean TOEIC Bridge difficulty = 0.54; Mean LOCAL test difficulty = 0.28

2.5.1 Interpretation of Fig. 2.1

As shown in Fig. 2.1 above, the discussion and the analyses can be summarized in the six points as follows:

- Pics, standing for pictures or photos, used in listening comprehension test did help Chinese test takers' listening comprehension. With photos, the listening comprehension part becomes much closer to the topics of real life and workplace. In this sense, listening comprehension tests without relevant photos or pictures only test the understanding of telephone conversations rather than the real life interactions because as we talk, mostly we could see each other, face to face.
- Understanding of the key words did help test takers' listening comprehension, too. This usually happens in the international communication setting, like a workplace, or in an office where the interlocutors are by no means native speaker of English, but understanding the key words and associating with the ideas in the picture or in the photo, test takers could understand and fulfil the international communication very well. Test takers who had the experiences of serving as a tour guide, working as intern in foreign office or family tutorial did the best.
- Voc, referring to frequent use of key words in English for daily life communication, helps listening comprehension. This happens to the test takers who often use English daily expressions on or off campus settings with their peers.
- Nearly all the subjects say that neither Chinese context on campus life nor grammar help learners in improving their listening comprehension. The typical examples are the phrases, sentence fragments they encountered in the dialogue. They tried to understand in the association with the specific context or setting but not with grammar.

- One additional yet important feedback concerning the listening comprehension feature is that in TOEIC Bridge Test, some speakers did not speak in standard American English; some are typically Korean's or Malaysian's. We think this is the very feature of TOEIC Bridge Test and should be maintained. In fact, this fully reflects the real life of workplace around the world because in any English speaking countries and even in the United States, not all the people speak English in the received standard American accent.
- Another feedback concerning the listening comprehension feature is that, to quote our students, the contents are much closer to the real life and workplace. To use the term of language testing, the materials are very authentic.

Therefore, in this sense, the listening comprehension part of TOEIC Bridge Test is a better measure of listening skills and could better verify the current level of the English proficiency of Chinese vocational students.

2.5.2 Interpretation of Fig. 2.2

Figure 2.2 above shows that the correlate coefficient is rather low, indicating almost no correlation existing and that big differences between the means and the difficulties of the two group scores are observed.

What we conducted here is a typical study of concurrent validity (See, e.g. Bachman and Palmer 1983). We met the two conditions here.

- Two sets of test items: listening comprehension parts of a local Chinese test and TOEIC Bridge test, of which TOEIC Bridge test is used as basal to measure the local Chinese one. The reason is simple. TOEIC Bridge test is an international test and recognized by hundreds of corporations worldwide.
- The formats of the two sets of test items are basically the same.

Here the only difference or improvement is that we mixed both test items together in one test paper. The results thus obtained show that TOEIC Bridge test fits Chinese students better and the score interpretation better indicates Chinese students' communicative competence in the real social and campus life. Evidences by comparison collected via questionnaires also indicate the possible reasons why the local parallel test is not very well recognized by educators and employers in China.

2.6 Concluding Remarks

This research, while focusing on the appropriateness of TOEIC Bridge Test to Chinese vocational students, reviews in some details the real situation regarding language testing in China, presents the research purpose and method supported with good analysis and thus can be concluded in three points as follows.

2.6.1 The Significances

The present study is the pioneer study conducted in language testing field in China which uses the real data to undertake the comparison of two tests being analyzed and illustrates some good features of TOEIC Bridge Test focusing on listening comprehension part. And it concludes that listening comprehension part of TOEIC Bridge test fits Chinese students better. This is not only based on the scores but also on the higher correlation between the TOEIC Bridge scores and the test takers' communicative competence in the real social and campus life.

The most significance of the present paper is to give warning message to our test counterparts and some authority concerned that the high stake of test exists and measures are to be taken to moderate it.

2.6.2 The Limitations

The two limitations for the present research are that the sample size is not big enough. And for the local test, the report of item analysis and test scoring generated by GiTEST shows that the discrimination indexes of many items are too low showing the test item are not very well calibrated or no pre-test was actually administered. The second limitation is that the similar comparison and analysis of reading comprehension should have been conducted. If it is, the overall picture of the appropriateness of the TOEIC Bridge test for Chinese students would be better silhouetted.

2.6.3 The Suggestions for Follow-Up Improvements

The author, by writing this paper, presents a universal yet serious problem not in all the vocational schools but also in most universities across China, i.e. the traditional practice of language testing under the guidance of classic testing theory has been largely neglected. The old generation of Chinese experts are either retired, or pushed aside, or give up. As the language testing is not attached great importance to, the current practice is that once a test is administered and the scores are reported in a school or to the school authority, the test papers are put aside. No one would think of anything else about the quality of test items and of the test paper in general. And this is partially the underlying reason(s) why the local test is not well recognized by both Chinese educators and employers. In this sense, the common practices like test item moderation, pre-test conduction, item analysis and test scoring and equating under the guidance of language testing theory implemented using Rasch-based or IRT-based software are to be promoted across China. More pre-conference workshops like IRT/CAT workshop, Rasch-, or IRT-based software

demonstration and interpretation are needed and more academic exchanges and cooperation in this line are to be encouraged. Only in this way could the test quality be improved and the test certificate be eventually recognized.

Acknowledgements We would like to thank the anonymous reviewer(s) and our foreign counterparts for suggestions that greatly helped improve the paper.

References

Ashmore, E., et al. (2009). *Official test-preparation guide*. Dalian: Dalian University of Technology Press.

Bachman, L. F., & Palmer, A. S. (1983). *Language testing in practice*. Oxford: Oxford University Press.

Hughes, A. (1989). *Testing for language teachers*. Cambridge/New York: Cambridge University Press.

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

Zhang, Q. (2011). *Towards better interaction between testing and teaching*. Keynote speaker at the 5th National TEFL/1st Mongolia TESOL Conference, Ulaanbaatar, Mongolia, Oct 7–9, 2011.

Siharay, S., et al. (2009). Appropriateness of the TOEIC Bridge test for students in three countries of South America. *Language Testing, 26*(4), 589–619.

Chapter 3
Validating the Model of Predictors of Academic Self-Handicapping Behavior

Hafsa Mzee Mwita, Mohamad Sahari Nordin, and Mohd Burhan Ibrahim

Abstract The main aim of the present study is to validate the model of predictors of self-handicapping behavior (POASH) on the data derived from undergraduate students in an ongoing co-curriculum compulsory course. The study adapted and extended the original theory of reciprocal interaction of emotion, cognition and behavior by adding self-handicapping behavior component. In so doing, this study assessed the direct and indirect effects of emotion, cognition and behavior via student engagement on self-handicapping behavior. The second purpose of the study is to evaluate gender and nationality status invariants of the causal structure of POASH. This cross-validation procedure determined whether gender and nationality status moderated the causal structure of the model, and thus the generality of POASH. The data was collected from two self-reported questionnaires administered to 790 undergraduates of an International Islamic University in Malaysia. A confirmatory three-step approach theory testing and development using Maximum Likelihood method was applied. The results of structured equation modeling supported the adequacy of POASH and the causal structure of POASH proved to be applicable to both genders and nationality statuses.

Keywords Self-Handicapping Behavior • Student Engagement • Structural Equation Modeling • Educational Psychology and Counseling

---

Major Part of Thesis to be submitted in Confirmatory with the Requirements For the degree of Doctor of PhilosophyInstitute of Education, Department of Counseling and Psychology International Islamic University Malaysia

H.M. Mwita (✉) • M.S. Nordin
Institute of Education, Department of Educational Psychology and Counseling, International Islamic University Malaysia, Selangor, Malaysia
e-mail: Hafsa.m.mwita@gmail.com; msahari@iium.edu.my

M.B. Ibrahim
Institute of Education, Department of Social Foundation and Educational Leadership, International Islamic University Malaysia, Selangor, Malaysia
e-mail: burhanibrahim@live.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_3,

3.1 Introduction

Academic self-concept is critical in the academic growth of the student because it has a direct effect on college performance, parents' & community expectations, student's future career, as well as his/her lifestyle and successes. Relative Emotive Behavioral therapy which is the parent of cognitive behavioral therapy is based on the assumption that cognitions, emotions and behaviors interact significantly and have a reciprocal cause-and-effect relationship (Corey 2013, p 267). This claim has been explained by Ellis (1993a) as the reciprocal interaction of emotion, cognition and behavior as demonstrated in Fig. 3.1, which has also been proven in a scientific study of Drevets and Raichle (1998), entitled "Reciprocal Suppression of Regional Cerebral Blood Flow during Emotional versus Higher Cognitive Processes". They claim that "the possibility that neutral activity in some cognitive-processing areas is suppressed during intense emotion states, which suggests mechanisms by which extreme fear or severe depression may interfere with cognitive performance" i.e. disengagement and self-handicapping behavior. In another scientific study on the relationship between emotion and cognition, Pessoa (2008), p 153 suggested that, "The cognitive control system guides behavior while maintaining goal-related information".

Thus, cognitive behavioral therapy is much more commonly used in the field of Academic clinical psychology (Jones and Butman 1991, p 145), which is therefore the most appropriate counseling theory in studying self-handicapping behavior of university students. Corey (1996 and 2013) suggested that REBT has consistently emphasized all of these three modalities and their interaction, thus qualifying it as integrative approach (Ellis 2001a, b, 2002, 2011; Ellis and Dryden 2007; Wolfe 2007). The present study extended this model by adding a self-handicapping behavior component. Therefore, survey has been conducted so as to identify undergraduates' self-concept on their emotional engagement, behavioural engagement, cognitive engagement, and self-handicapping behaviour. This is the first study to be conducted on the reciprocal interaction of emotion, behaviour and cognition as predictors of self-handicapping behaviour of Muslim University Students thus, no previous study to compare with.

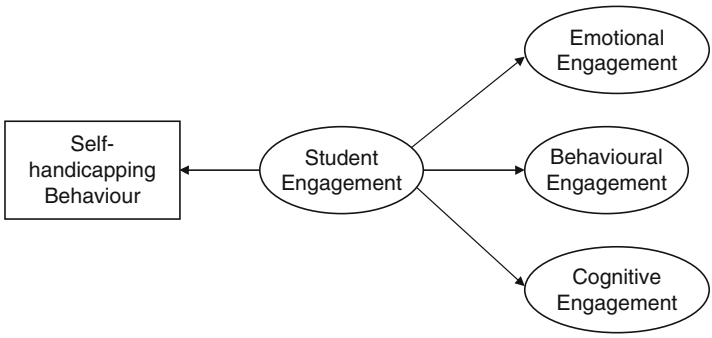


Fig. 3.1 Theoretical model of POASH (Mwita et al. 2013)

The design of a four construct prediction model of self-handicapping behaviour (POASH), which is also a regression model, is based on the *Rational Emotive Behavioral Therapy of Albert Elis* (1993) and the study of *Predictors of Physical Spouse/Intimate Violence in Chinese American Families* (Yick 2000), whereby his study followed previous studies that have tested for the validity of second order factorial structure e.g. Dobash and Dobash (1979) and Straus (1980). Components of the model are based on learned helplessness theory of Seligman and Maier (1967); a study on Self-Handicappers: Individual Differences in the Preference for Anticipatory Self-Protective Acts (Rhodewalt 1990); and on a study on Quantifying School Engagement by the National Center for School Engagement (2006) in Colorado, USA. Therefore, this model which has been designed by the author and has not undergone previous test thus, needed a thorough test before its use. Therefore, by application of SPSS 16, all four scales were analyzed separately through Component Factor Analysis, before implementing Confirmatory Factor Analysis through Amos 16. Thereafter, group analysis of measurement models took place and finally the analysis of the structure model including invariant analysis.

Thus, in this study researchers assumed a negative influence of all the three constructs of student engagement on academic self-handicapping behavior. This led into the formation of the theoretical model of this study (Fig. 3.1), which hypothesized a significant reciprocal interaction between emotional engagement, cognitive engagement and behavioral engagement which would function as predictors of self-handicapping behavior of undergraduate students.

3.2 Methodology

3.2.1 Introduction

This is a theoretical study which deals with model building, assessment and evaluation through structural equation modeling. It involves a confirmatory three-step approach theory testing and development using Maximum Likelihood method was applied. Firstly, the four constructs of the hypothesized model of POASH were assessed through PCA and CFA, whereby the three student engagement constructs (emotion, behavior and cognitive) proved to be fit and reliable while the SHB construct proved to be fit but unreliable due to low AVE and r^2 therefore, we rejected it and replaced it with the composite score. Secondly, the individual constructs were assessed as a group of constructs by embedding the three (emotional engagement, behavioral engagement and cognitive engagement) university student engagement constructs together as a measurement model of USE before assessing its fitness in the form of first and second order measurement models. Thirdly, the structure model of POASH was built and its goodness of fit was examined before cross-validating it by assessing the moderating effects of gender and nationality status on the structured model of POASH.

The hypothesized models were estimated using the covariance matrix derived from the data. Thus, the estimation procedure satisfied the underlying statistical distribution theory, and yielded estimates of the desirable properties. The study adopted maximum likelihood in generating estimates of the structured model of POASH. After the estimation of the model we applied a set of conventionally accepted criteria for deciding on the constituents of a good fit model by assessing the: (a) consistency of the hypothesized model with the empirical data, (b) reasonableness of the estimates (c) the proportion of the variance of the dependent variables accounted for by the exogenous variables.

3.2.2 Sample

From our target population of 1,032 students, only 832 responded and only 790 students followed the instructions and filled in the survey report correctly and completely thus, 42 samples were discarded due to either incorrectly filled in questionnaire or partially filled or not filled. Therefore, the total sample included in the final analysis is 790 undergraduate students, 272 (34.4 %) are male and female students are 518 (65.6 %). These percentages almost resemble the overall admission of IIUM which is 40 % male students and 60 % female students. Age range of the respondents is between 18 and 29 years whereby the majority are 20 years old (75.1 %) followed by 21 year olds (14 %), 19 years (3.8 %), 22 years (3.4 %), 23 years (4 %) and the rest are less than 1 %. This age range is very appropriate in studying self-handicapping behavior which according to most of the previous studies reported that self-handicapping behavior mostly occurs between the ages of 18 and 25 years.

3.2.3 Data Screening of SEQ

Descriptive statistics of all 44 items of Student Engagement Questionnaire (SEQ, 2011) from the whole sample (N = 790) was done through SPSS 16. The score of means are assessed from 7-points Likert scale ranging and results show a range of 3.88 to 6.78 and the standard deviations of 0.86 to 2.09. The statistic value (z) of the skewness and kurtosis fell below the threshold point of the skewness (−3 to +3) and kurtosis (−10 to +10) as noted by Kline (2011), except for item ce1 (skewness = −3.475) which was later-on removed. The reliability estimates for internal consistency for 45 items of the three scales (N = 790) are: emotional engagement – Cr. 0.88, behavioural engagement – Cr. 0.81 and cognitive engagement – Cr. 0.842 from a scale of 1–7.

The mean score is 5.28; the minimum and maximum scores range from 3.88 to 6.784; and the Standard deviation is from .591 to 1.96. The statistical value (z) of the skewness and kurtosis fell below the threshold point of the skewness (−3 to +3) and kurtosis –(10 to +10) as noted by Kline (2011), all are within the acceptable limits

except for item EE1 which was later on removed. Thus based on the result of descriptive statistics, SEQ was considered to be a highly reliable instrument. And which is inconsistent with the findings of previous studies as reported by Finlay (2006).

3.2.4 Data Screening of SHQ

Descriptive statistics of all 20 items of Self-Handicapping Questionnaire (SHQ, 2011) from the whole sample (N = 790) was done through SPSS 16. The score means were assessed from 7-points Likert scale result indicate a range of 1.65 to 5.10 and the standard deviations 1.37 to 1.99. In reference to Kline (2011), the statistic value (z) of the skewness and kurtosis fell below the threshold point of the skewness (−3 to +3) and kurtosis (−10 to +10); the reliability estimate for internal consistency for 20 items of the self-handicapping questionnaire (N = 790) is 0.78, thus considered to be a reliable instrument.

3.3 Analysis of the Measurement Models

3.3.1 Factorial Validity of the Measurement Model of USE

Confirmatory Factor Analysis was applied in-order to ensure the maximum results to which the observed items are to be generated by the underlying latent constructs, which finally provided the links between the latent variables and observed variables (Byrne 2010).

Results of the descriptive statistics indicated the mean of 5.28 from a scale of 1–7; the minimum and maximum scores range from 3.88 to 6.784; and the Standard deviation is from 0.591 to 1.96. The statistical values (z) of skewness fell below the threshold point of −3 to +3 and kurtosis fell below −10 and +10 thus, all were within the acceptable limits except for item CE11 with kurtosis of 14.63 which has been eliminated from further analysis. Outliers were determined by observing the Mahalanobis distance which is the farthest point from the centroid, and all items that fell under the high Mahalanobis d-squared with both P1 and P2 equal to 0.000 were considered as outliers and therefore removed. Factor Loadings of individual constructs and component fit measures have been examined to check whether any construct would be rejected, but none was rejected because all the factor loadings are above 0.5 and AVE of 0.5 which proves the hypothesis of question one which says each factor substantially influences its targeted indicators, each of which accounts for more than 50 % of the variance explained (However, the hypothesis was later tested against SHB). In evaluating the alternative models further reduction of items took place. Consideration was mostly given to the overall fit measures, based on a

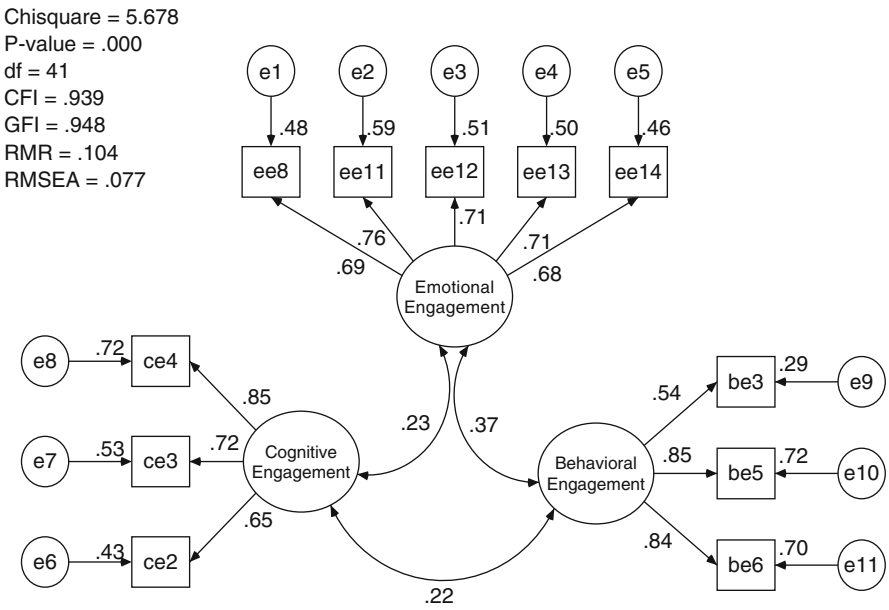


Fig. 3.2 Hypothesized first order measurement model of USE

number of fit indices namely – Normed chi-square (cmindf), Significance (*P*) degree of freedom *(df)*, Comparative Fit Index (CFI), Goodness of Fit Index (GFI), Root Mean Square Residual (RMR), Root Mean Square Error of Approximation (RMSEA). The component fit measures came from parameters estimates (Bollen 1989), which included the squared multiple correlation (R^2) for each pair of relationship. Therefore, the accepted constructs in terms of overall fit and component fit were then knitted together to form the measurement model of USE (Fig. 8). This has also been suggested by Hall, Snell and Foust (1999) who stated that "For theoretical and empirical reasons researchers may combine item-level responses into aggregate item parcels to use as indicators in a structural equation modelling nest". By using SEM, the researcher examined the relationships between the observed variables and the latent variables in the analysis of the full measurement model of USE.

3.3.2 CFA and Results of the Measurement Model of USE

The measurement model (Fig. 3.2) is a first order confirmatory factor analysis model designed to test the multidimensionality of USE model i.e. to test the hypothesis that the multidimensionality construct of USE is composed of inter-related constructs of emotional engagement, behaviour engagement and cognitive engagement. Offending estimates were searched for, fortunately the measurement model indicated absence of negative error variances i.e. absence of standardized coefficients exceeding 1.00, extreme values of standard errors and the residuals

greater than 2.58. After a series of CFA, 11 items out of 13 items have been selected thus; the hypothesized model consists of the three inter-correlated factors with 11 observed variables (ee8, ee11, ee12, ee13, ee14, be3, be5, be6, ce2, ce3, and ce4). Each observed variable was hypothesized to load onto one factor only.

The researchers assessed the hypothesized model to determine to what extent the model fits the sample data. Almost all indicators were found to have good significant loadings with respect to model adequacy as a whole: the measurement of normed chi-square = 5.678, *df* = 41, CFI = 0.939, GFI = 0.948, RMR = 0.104, RMSEA = 0.077. Feasibility of the individual parameters of the factor loading was estimated, results demonstrated in Fig. 3.2, indicates a range of factor loadings from 0.54 (be3) to 0.85 (be5 and ce4). Thus, the requirement for convergent validity of $\geq$0.5 and not exceeding 1 has been fulfilled. The observed variables which measures a common underlying factor are all found to be statistically significant i.e. Critical Ratio (CR) >1.96, while the Standard Error (SE) range from 0.061 to 0.151, the variances of error terms range from 0.329 to 1.435 and factor variances ranges from 0.469 to 1.022 are all within the significant range of ± 2.58 (Kline 2011).

According to the results of Maximum Likelihood Parameter Estimates, squared multiple correlation shows the factor of behavioural engagement is explained by 71.7 % variance of be5, followed by 70.2 % variance of be6, and 29.2 % associated with variance of be3. Cognitive engagement factor is explained by 71.8 % variance of ce4, followed by 52.5 % variance of ce3, followed by 42.8 % variance of ce2. Emotional engagement construct is explained by 58.5 % variance of ee11, followed by 50.5 % variance of ee12, followed by 50.3 % variance of ee13, followed by 47.9 % variance of ee8 and 46.1 % variance of ee14. These results indicate that almost all the loadings are statistically significant good predictors (46.1–71.7 %) except one predictor be3 which is of average significance percentage of 29.2 %. The latent factor correlations are significant and positively correlated with r = 0.366 (behavioral and emotional engagement), r = 0.218 (behavior and cognitive engagement), r = 0.234 (cognitive and emotional engagement). The result of correlation among three latent factors of USE model indicates no correlation of above 0.85 and none of bellow Critical Ratio of >1.96 i.e. none of the values is above 0.01 significance. This supports the discriminant validity upon which factors are independent and yet they are moderately correlated.

Convergent validity which is referred to a set of variables (items) that presume to measure a construct (Kline 2005) and discriminant validity which refers to the extent in which a construct is truly distinct from other constructs (Byrne 2010; Kline 2011), was carried-out in the process of assessing the set of variables within the three factors which represents the student engagement scales (emotional, behavioral and cognitive engagement). Despite of having their significant loadings, the student engagement items vary significantly as to the degree to which they explain the factor. The factor loadings are all within and above their expected limits.

Average Variance Extracted (AVE) for each construct was compared against the square of correlation between the items within each factor and all AVEs are >0.5. According to Fornell and Lacker (1981), AVE $\geq$0.5 indicates high convergent validity; and according to Hair et al. (2010), factor loadings $\geq$0.5 indicate high

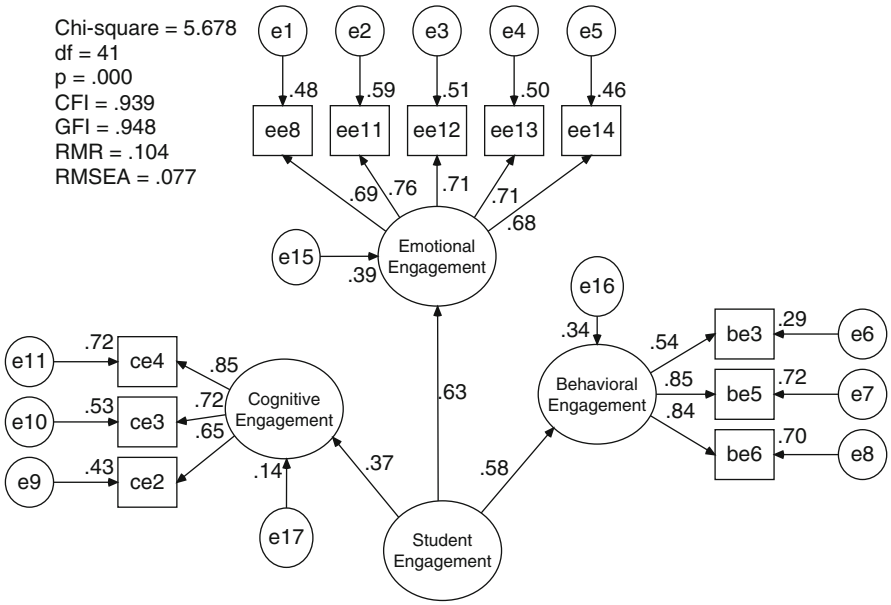


Fig. 3.3 Hypothesized second order factor of the measurement model of USE

convergent validity i.e. above 50 %, thus, all our three constructs are considered to have a high convergent and discriminant validity and therefore all three factors have been retained. Results also indicated the values of the residual co-variances which fell below the threshold point of Multicolinearity of <2.58 (Hair et al. 2010) therefore; the model is accepted even without the re-specification of the Modification Indexes (MI).

3.3.3 CFA and Results of the Second Order Measurement Model of USE

The second order CFA model was designed according to the hypothesized model of POASH whereby the model tested in the present application hypothesized a priori that, responses to the Student Engagement can be explained by three first order-factors (emotional engagement, behavioural engagement and cognitive engagement); each item has a nonzero loading on the first-order factor it was designed to measure, and a zero loadings on the other two first-order factors; error terms associated with each item are uncorrelated; co-variation among the three first-order factors is explained fully by their regression on the second order factor. By using the maximum likelihood procedure of the confirmatory factor analysis the validity of second order factor was tested after the first order factor of the model of USE. Two of the first order factors are measured by three items and the third factor is measured by five items. Each item is loading on its own factor only, as indicated in Fig. 3.3.

Results indicate that the hypothesized first and second order measurement models provide a good explanation of the model of USE in the current study. With its three inter-related factors (emotional engagement, behavior engagement and cognitive engagement) and 11 measured variables, this model supports the hypothesis that the measurement model of USE is a multidimensional construct consisting of emotional engagement, behavior engagement and cognitive engagement. The overall ft of the model is adequate as depicted in the model and as explained in the results of the first order measurement model.

All factor loadings define their respective factors, and factor correlations are of moderate size while representing their distinct constructs. The co-variation among the three first-order factors is explained fully by their regression on the second order factor. Therefore this result affirms the two hypothesis of research question one whereby the first one affirms that each factor substantially influences its target indicators; each of which accounts for more than 50 % of the variance explained and the second one which affirms that the hypothesized measurement model of USE adequately fits the data. Moreover, it affirms the single hypotheses of research question two which claims for the occurrence of a significant inter-relationship between emotional, cognitive, and behavioural, engagement of undergraduate students.

3.3.4 CFA and Results of the Measurement Model of SHB

The final 15 items which have been extracted after the PCA of self-handicapping behavior questionnaire, were used in executing the CFA of SHB construct. According to the results of the fit indices which were obtained after a series of measurement model comparison which were assessed according to their fit indices, self-handicapping behaviour model proved to be perfectly fit after obtaining a chi-square = 4.823, df = 2, P-value = 0.008, CFI = 0.972, GFI = 0.994, RMR = 0.075 and RMSEA = 0.070. However, the result of both AVE and multiple square correlations were less than the acceptable level of 0.5 therefore, the model was rejected based on the fact that it failed to prove the first hypothesis of research question one. Thus, according to these findings, we decided to use the composite score of SHB scale as a replacement of SHB construct and hence, the final design of POASH was opted by using the short SHQ-15 which the researchers nested it to the measurement model of USE. The reliability statistics of the 15 items of the Short SHQ-15 for the total population of 790 undergraduates indicate a standard Cronbach's Alpha of 0.696. From a scale of 1–7, the mean is 4.045; the minimum and maximum scores range from 2.087 to 5.100; and the Standard deviation is from 1.369 to 1.98. The statistical values (z) of skewness fell below the threshold point of −3 to +3 (Kline 2011), and kurtosis fell below −10 and +10 thus all fifteen items are good items which are within the acceptable limits.

3.4 Analysis of the Main Study

3.4.1 Adequacy of Causal Structure of the Model of POASH

Figure 3.4 summarizes the results of the structural equation modeling of the POASH model. The confirmatory modeling yielded consistency of the hypothesized causal relationships with the data. (Nomed Chi-square = 5.4, RMSEA = .075, RMR = .36, GFI = .94 and CFI = .95). All these fit indices satisfy their critical cut-score result thus, indicating a good fitting model of POASH. The parameter estimates of the hypothesized structured model are free from offending values. All path coefficients of the causal structure are statistically significant at 0.005 levels, and are of practical importance, since the smallest value of the standardized path coefficient is 0.2.

Results indicate that behavioral engagement is relatively more influential than both emotional and cognitive engagement in directly influencing SHB via student

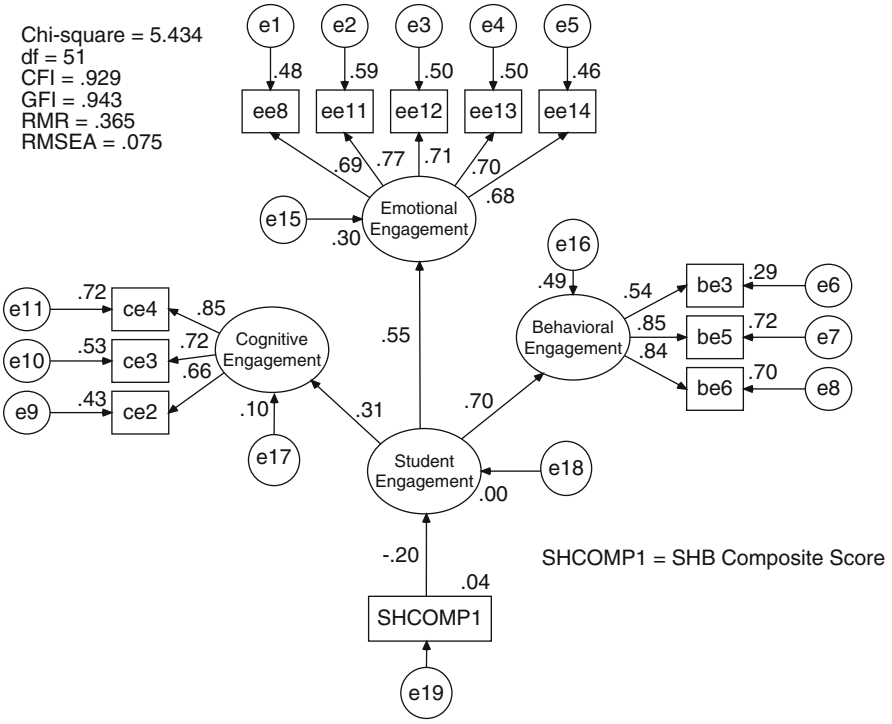


Fig. 3.4 Standardized coefficients of the hypothesized model of POASH. *Note: emotional engagement, behavioral engagement and cognitive engagement are exogenous latent variables (IVs), while the composite score of self-handicapping behavior is endogenous manifest variable (DV). The three exogenous variables (EE, BE and CE) significantly influence the SHB endogenous variable* via *student engagement latent variable which directly and negatively influences SHB*

engagement, while emotional engagement influences SHB more than the cognitive engagement does. The influence of the three construct through a latent variable of student engagement shows that self-handicapping behavior is explained by 70 % variance of behavioral engagement, 55 % variance of emotional engagement and 31 % variance of cognitive engagement.

Moreover, the analysis reveals that collectively the variability of three exogenous variables directly explain 20 % of self-handicapping behavior, as well as the negative influence of student engagement on self-handicapping behavior (Fig. 3.4). This also explains the reciprocal interaction of emotion, behavior and cognition; all three constructs need to support each other in a triple action. This finding supports the previous findings as stated by Corey (2013, p 267) who reported that "Relative Emotive Behavioral Therapy (REBT) has consistently emphasized all three of these modalities and their interaction, thus qualifying it as integrative approach (Ellis 2001a, b, 2002, 2011; Ellis and Dryden 2007; Wolfe 2007)". In total, the results answered the third research question and provided the support for the four hypotheses of research question three.

3.4.2 The Test of Equivalence of the Structure Model Across Groups

3.4.2.1 Gender Invariance of the POASH Model

Both gender groups were analyzed simultaneously in order to identify if the comparisons between the structural models can be interpreted similarly. The fitness of the baseline models was the same for both male and female when analyzed simultaneously and freely estimated without being constrained. The freely estimated model became the bench mark for the comparison of the fit statistics.

Results indicated a fit model without being constrained, whereby the fit statistics revealed CFI = 0.925, GFI = 0.932, RMR = 0.369, RMSEA = 0.054. Normed chi-square = 3.323 with a degree of freedom of 102 (see Table 4.2 and Fig. 3.5). Convergent validity can be explained by observing the factor loadings of above 0.5 in both models although they are not having a similar weight which explains the divergent validity. Similarly, divergent validity is explained by differences in the multiple square correlations of the two models.

However, the gender model was then constrained and the result (Table 3.1) confirms our finding and proves our hypothesis. The chi-square difference between the constrained and unconstrained model is only 3.279 and the difference between the two degrees of freedom is only 3, which is an insignificant figure at the significance level of 0.005. The results also proves that the model is significant with a P-value of 0.351 while the chi-square thresholds also prove significance at 90 % confidence, 95 % confidence and 99 % confidence.

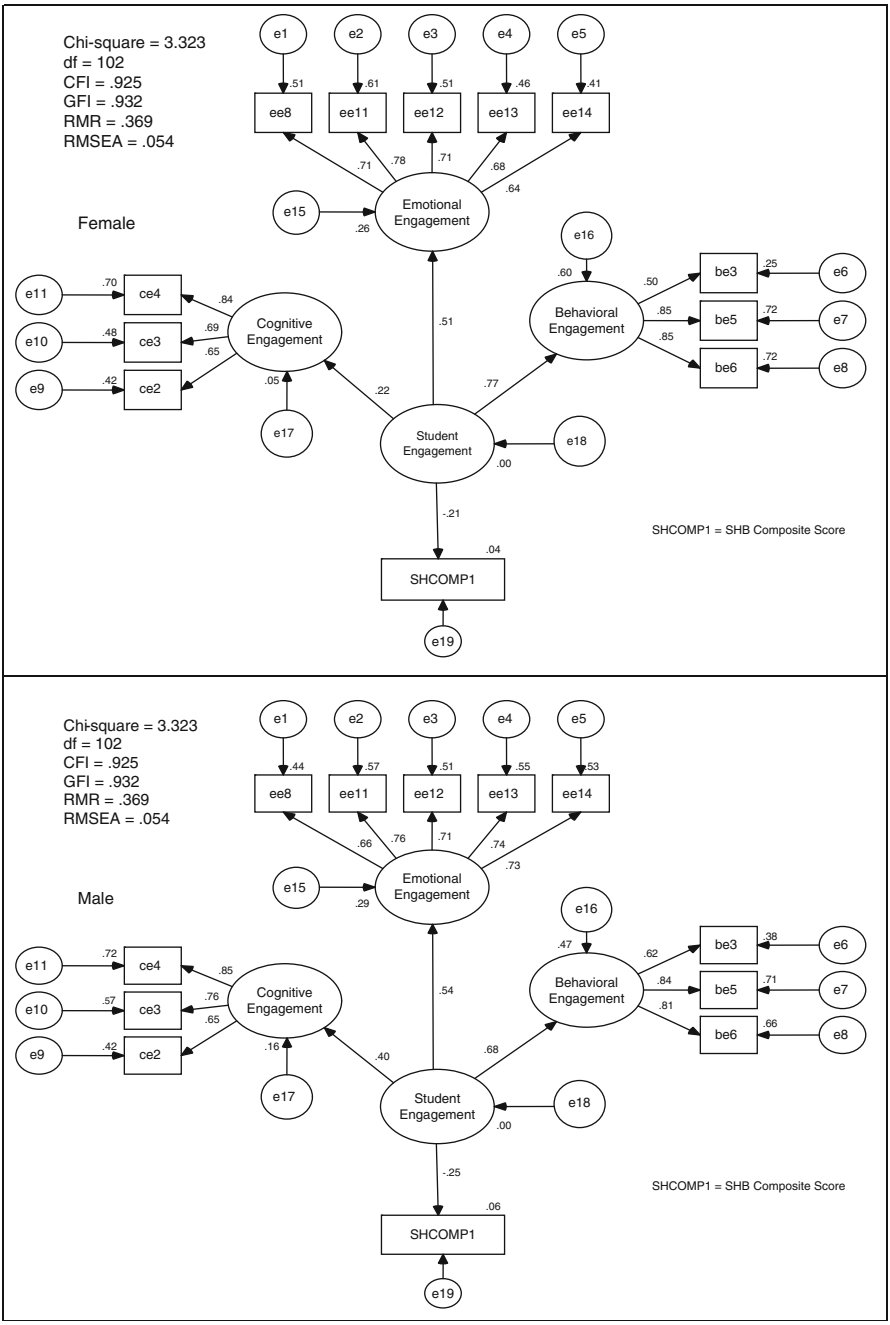


Fig. 3.5 Simultaneous analysis of gender invariance of the POASH model

Table 3.1 Constraining the gender invariance model

	Chi-square	df	P-value	Invariant?
Overall model				
Unconstrained	338.93	102		
Fully constrained	342.209	105		
Difference	3.279	3	0.35	Yes
Chi-square thresholds				
90 % confidence	341.64	103		
Difference	2.71	1	0.10	
95 % confidence	342.77	103		
Difference	3.84	1	0.05	
99 % confidence	345.56	103		
Difference	6.63	1	0.01	

3.4.2.2 Nationality Status Invariance of the POASH Model

In analyzing the nationality status invariance of the structured model of POASH, a second pair of baseline measurement model was established for both national and international groups of students separately so as to select the best fitting model. Once again the model was scrutinized and selection of the best fitting model was done according to parsimony. Fit indices were assessed to confirm the status of the replication of both models in a simultaneous testing. Both baseline models of nationality status invariance proved to be the same before constraining (Fig. 3.6), whereby the fit statistics revealed CFI = 0.926, GFI = 0.910, RMR = 0.798, RMSEA = 0.056. The normed chi-square is 2.114 with a degree of freedom 102. Although the P-values shows insignificant result in the model but it proved to have a significant value of 0.788 when the baseline model of nationality status was constrained by using online package statistical tool.

The model also indicates that national students self-handicap more than the international students where 25 % variance of student engagement directly explains self handicapping behavior of national undergraduates while an insignificant figure of 12 % of student engagement directly explains self handicapping behavior of international students. However, the variability of behavior engagement explains the highest in international students (69 %) when compared to national students (68 %). Followed by emotional engagement which is also explained higher (68 %) by the international students when compared to national students (54 %). However, variability of cognitive engagement of national students explains SHB higher (40 %) when compared to the variability of cognitive engagement of international students (38 %).

To sum-up, both Nationality Status and gender, do not moderate the structure model. This fact is verified by the observation of the identical values of fit statistics for the two pairs of models (Figs. 3.5 and 3.6). Moreover, the result of online statics tool (Tables 3.1 and 3.2) further confirmed the findings.

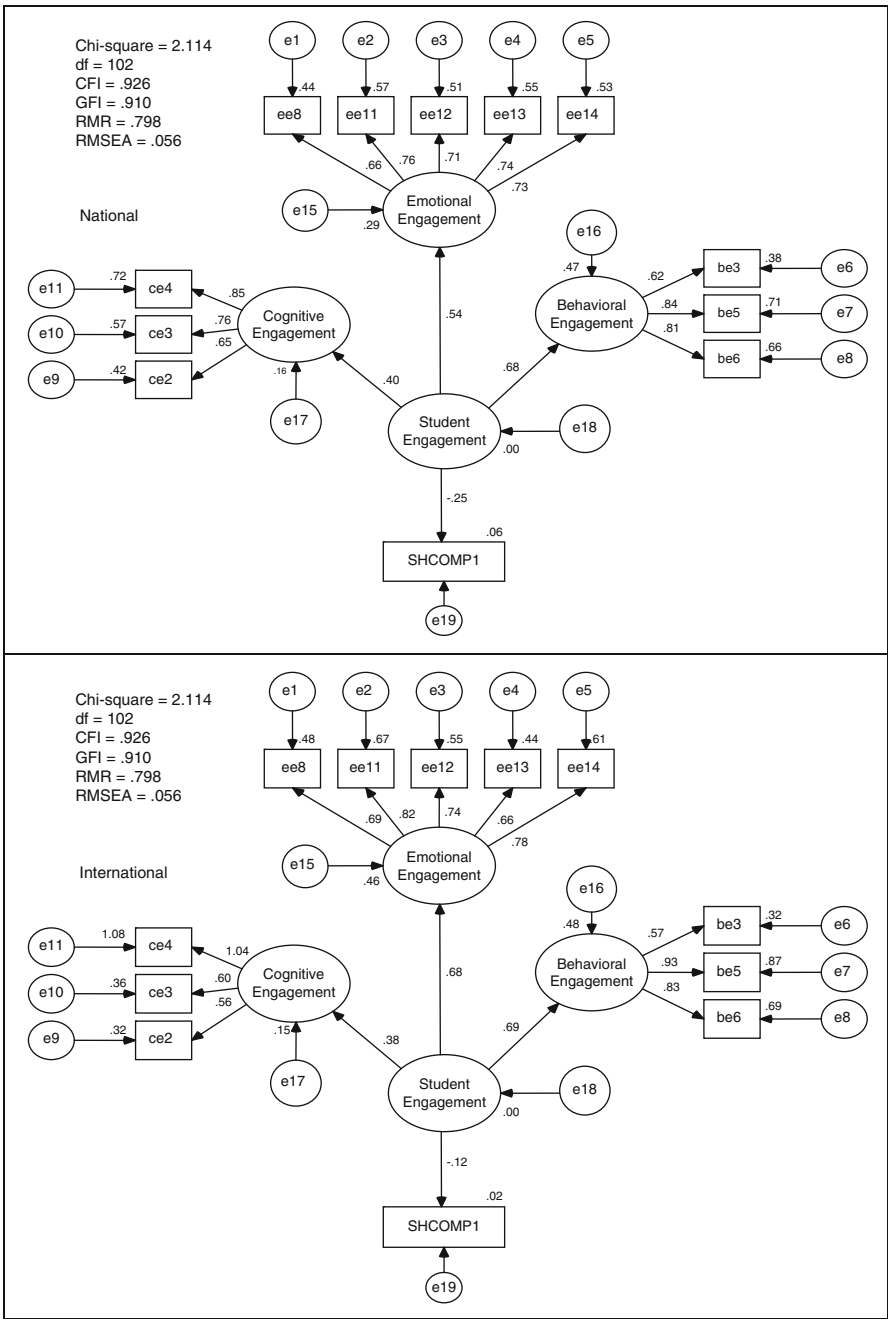


Fig. 3.6 Simultaneous analysis of nationality status invariance of the POASH model

Table 3.2 Constraining the nationality status invariance model

	Chi-square	df	P-value	Invariant?
Overall model				
Unconstrained	215.608	102		
Fully constrained	216.661	105		
Difference	1.053	3	0.79	Yes
Chi-square thresholds				
90 % confidence	218.31	103		
Difference	2.71	1	0.10	
95 % confidence	219.45	103		
Difference	3.84	1	0.05	
99 % confidence	222.24	103		
Difference	6.63	1	0.01	

3.4.3 Summary of the Analysis of the Main Study

In the main analysis of this study, the three step approach which involves the analysis of measurement models, analysis of structure model and estimating the structural invariance across groups of gender and status of nationality has been explained in detail. The first research question led the researcher in to analyzing the measurement model of USE and the measurement model of SHB in order to determine the significant influence of each factor on its related construct. Results indicated that each factor among the three factors of the measurement model of USE substantially influences its target indicator by more than 50 % variance. However, the single construct of SHB measurement model could be explained by less than 50 % variance although the fit indices indicated that it was a very good fitting model. Thus, the model was rejected and replaced by the composite score of 15 good items from the initial CFA. The answer to the second research question proved that the measurement model of USE fits the data. It also proves the reciprocal interaction of emotion, behavior, and cognition, which is the assumption of the REBT of Albert Ellis. Answer to the third research question proved the significant negative influence of student engagement on SHB. In answering the third question it was revealed that male students have a higher self-handicapping tendency when compared to female students. Finally, the answers to question four proves that the structured model of POASH fits the data and can be replicated across independent samples drawn from the same population, Thereby, proving our entire hypotheses to be true. Furthermore, in answering question four, it was revealed that male students self-handicap more than the female students, and national students self-handicap more than the international students.

3.5 Discussions

3.5.1 Discussion

Findings of the present study have expanded the existing body of knowledge on the reciprocal interaction theory of emotion, behavior and cognition in several ways: **First,** the results substantiated the psychometric adequacy of the measure of student engagement model as well as the psychometric adequacy of the measure of the structured model of predictors of academic self-handicapping behavior. Both Measures seems to be sufficient to represent the measurement tools of assessing student engagement and student's academic self-handicapping behavior. **Second,** the results validated the good fit of the model of predictors of academic self-handicapping behavior (POASH). The results supported the efficacy of the original model of reciprocal interaction of emotion, behavior and cognition of which posits that cognitions, emotions, and behaviors interact significantly and have a reciprocal cause and effect relationship. In addition the results are congruent results of previous studies (Ellis 2001a, b, 2002, 2011; Ellis and Dryden 2007; Wolfe 2007) which also found the significant relationship of emotion, cognition and behavior. **Third,** the structure equation model of POASH proves to be fit and reliable and can be applied in analyzing data of different groups because it will produce equal results. **Fourth,** The model of nationality status invariance also indicates that national students self-handicap more than the international students where 25 % variance of student engagement explains self handicapping behavior of male undergraduates, while insignificant figure of only 12 % of international students' engagement explains self handicapping behavior of international students. This finding is explained by Status characteristic theory of Berger, Cohen, and Zelditch (1996) which states that members of a group form expectations about each other's competence to contribute to group goals based on their status characteristics. In addition, Podolny (1993) explained that Individuals expected to contribute more are more highly valued by the group, held in higher esteem. In our situation, national students tend to be more highly valued and esteemed than international students therefore; this might be the reason why national students seem to self-handicap more than internal students so as to protect expectations for personal competence. **Fifth,** International students' behavior engagement and national students' behavior are almost the same i.e. 69 % of behavior engagement of international undergraduates explains SHB while 68 % of behavior engagement of nationals' explains their SHB. This is probably due to the same characteristics of young adults no matter which nationality they belong to. This is supported by Arnett (2007) who reported that: "One claim made frequently about emerging adults is that they are miserable lot, wracked with anxiety & unhappiness, intimidated to the point of paralysis about the grim prospects for entering the adult world. According to this view, the years from age 18 to 25 are dark and dreary period of life course. Emerging adults are typically confused and glum and overwhelmed by what the world seems to require from them". **Six,** results indicate that emotional engagement of international students

explains 68 % of SHB while only 54 % emotional engagement of national students explains their SHB. This is supported by the finding of Newman and Wadas (1997), who claim that the differences in the use of self-handicaps between persons with stable versus unstable self-esteem were much stronger than between persons with generally high versus low self-esteem. **Seven,** variance of cognitive engagement of national students 40 % has a very small difference with that of international students' which is 38 %. The low percentages and the small differences between them might be due to difficulty of interpreting and explaining cognitive engagement.

3.5.2 Implication of the Study

Findings of this proposed study will be noteworthy because previous research of this nature has not been conducted on Muslim University Students. More specifically the findings of this study will represent IIUM. The information on how self-handicapping behaviors varies between the three categories of engagement, has both theoretical and practical implications for educators, counselors, psychologists including student counselors and psychologists who are interested in understanding student's behavior and appropriate actions to be taken. The results of this study did not only contribute to the literature and researches done on, self-handicapping behavior, student engagement and on the reciprocal inter-action of emotion, behavior and cognition but, also helped the counselor/researcher with the identification and reduction of self-handicapping behaviors among the students while carrying-out this study. Thus, the person who benefitted the most is the student who is able to understand his/her problems and also know that he/she can be helped through counseling and guidance service. Individual Counseling was carried out according to individual requests. Study results shall be presented to Co-Curricula Activity Centre (CCAC), and Counseling Unit of IIUM, for appropriate actions. The first part of this study has been presented at the Pacific-Rim Objective Measurement Symposium organized by Jiaxing University in China (August 6th–9th 2012).

3.5.3 Conclusion and Recommendations

According to the American School Counsellor Association (ASCA) Counselling Model, the role of a counsellor is to promote academic, career, personal, and social life of every student. Therefore University Counsellors are supposed to have structured developmental lessons which are connected with academic areas; work with students and families to help all students develop personal goals and future plans; meet immediate student needs through crisis counselling, referrals and follow-up, target activities and maintaining as well as enhancing the educational environment and school climate. Since student engagement is self manageable,

university counseling unit may consider self enhancement programs for the students and staff development programs for lecturers and facilitators on identifying and rectifying academic self-handicapping behavior and academic disengagement of students. Counselors, psychologists and researchers (including students) may be urged to conduct more studies on self-handicapping behavior and student engagement as it appears to be very common among the young adults of all genders and cultures.

References

Abu Samah, B. (2012). *Structural equation modeling (SEM): Using AMOS*. Department of Professional Development and Continuing Education. Faculty of Educational Studies. Universiti Putra Malaysia. Serdang, Malaysia.

Al-Qayyim, I. I. (2003). *Healing with the medicine of the prophet* (pp. 177–180). Edited by: Abdul Rahman Abdullah Formarly Manderola R. J. Riyadh Saudi Arabia: Darussalaam Publishers & Distributers.

Al-Qayyim, I. I. (2004). *Alfawâ'id: A collection of wise sayings* (Translation of Shafiq bin Ibrahim bin Aly Al-Azdy Al-Balkhy, p. 295). Al-Mansura: Umm Al-Qura for Translation, Publishing and Distributions.

American Psychiatric Association (2000). *Diagnostic and scientific manual of mental disorders* (4th ed., Text Revision). Washington, DC: Library of Congress Cataloging-in Publication Data.

Arnett, J. J. (2007). Emerging adulthood, a 21st century theory: A Rejoinder to Henry and Kloep. *Society for Research in Child Development, 1*(2), 80–82.

Baba, S. (2010). *Lecture notes on Islamization of knowledge*. International Islamic University Malaysia. Gombak, Malaysia.

Badri, M. B. (2009). *Interview session between a PhD counseling students and Malik Badri*. Institute of Education, IIUM, KL Malaysia.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606.

Berger, J. W., Bochow, T. W., Talamo, J. H., & D'Amico, D. J. (1996). Measurement and modelling of thermal transients during Er:YAG laser irradiation of vitreous. *Lasers in Surgery and Medicine, 19*(4), 388–396.

Berger, J., Cohen B. P., & Zelditch, M. (1972) In Berger J., & Zelditch M. (2002). *New directions in contemporary sociological theory* (p. 69). Lanham: Rowman & Littlefield Publishers, Inc.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons, Inc.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230–258.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications and programming* (2nd ed., pp. 5, 53–160). New York: Routledge Taylor & Francis Group.

Corey, G. (1996). *Theories and practice of counseling and psychotherapy* (5th ed., pp. 134–135). Pacific Grove, CA: Brooks/Cole Publishing Company, 317–357.

Corey, G. (2013). *Theory and practice of counseling therapy* (9th ed., pp. 263–300). Belmont: Brooks/Cole Publishing Company.

Counsel for the Advancement of Standards in Higher Education (CAS). (1999). *The role of counseling programs: CAS standards contextual statement*. Washington, DC: CAS.

Creswell, J. W. (2005a). *Educational research planning, conducting, and evaluating quantitative and qualitative research* (pp. 509–515). Upper Saddle River: Pearson Education, Inc.

Creswell, J. W. (2005b). Practical action research. In *Educational research – planning, conducting, and evaluating quantitative and qualitative research* (pp. 41, 324–349, 354, 552–555). Upper Saddle River: Pearson Education, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. In Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4).

Dawry, M. (2006). *Counseling and psychotherapy with Arabs and Muslims: A culturally sensitive approach* (p. 58). New York: Library of Congress Cataloging in Publication Data.

Dobash, R. E., & Dobash, R. P. (1979). *Violence against wives*. New York: The Free Press.

Dolcos, F., Iordan, A. D., & Dolcos, S. (2011). Neural correlates of emotion-cognition interactions: A review of evidence from brain imaging investigations. *Journal of Cognitive Psychology, 23*(6), 669–694.

Drevets, W., & Raichle, M. (1998). Reciprocal cerebral blood flow during emotional versus higher cognitive processes. *Cognitive and Emotion, 12*(3), 353–385.

Ellis, A. (1993a) In Corey, G. (2013 & 1996). *Theories and practice of counseling and psychotherapy* (9th ed., pp. 267–286 and 5th ed., pp. 317–357). Pacific Grove, CA: Brooks/Cole Publishing Company.

Ellis, A. (1993b) In Corey, G. (1996). *Theory and practice of counseling therapy* (5th ed., pp. 317–359). Pacific Grove, CA: Brooks/Cole Publishing Company.

Ellis, A. (2001a). *Feeling better, getting better, and staying better*. Atascadero, CA: Impact.

Ellis, A. (2001b). *Overcoming destructive beliefs, feelings, and behaviors*. Amherst, NY: Prometheus Books.

Ellis, A. (2002). *Overcoming resistance: A rational emotive behavior therapy integrated approach* (2nd ed.). New York: Springer.

Ellis, A. (2011). Rational emotive behavioral therapy. In R. Corsini & D. Wedding (Eds.), *Current psychotherapy*. Washington, DC: American Psychological Association.

Ellis, A., & Dryden, W. (2007). *The practice of rationale behavior therapy* (2nd ed.). New York: Springer.

Erickson, E. H. (1963). Childhood and society. In Corey, G. (2013) *Theory and practice of counseling therapy* (5th ed., pp. 63–66). New York: Brooks/Cole Publishing Company.

Erickson, E. H. (1968). Identity: Youth and crisis. Norton, New York, 1968. *Science, 161*(3838), 257–258.

Finlay, K. A. (2006). *Quantifying school engagement: Research report*. National Center for School Engagement. c/o Colorado Foundation for Families and Children. Denver.

Fornel, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Fredricks, J., Blumenfel, P., & Paris, A. (2004). School engagement: Potential concept, state of evidence. *Review of Educational Research, 74*, 59–109.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS for windows and Mackintosh: Analyzing and understanding data* (4th ed., pp. 273–297). Upper Saddle River: Pearson Prentice Hall.

Hair, J. B., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Hall, R. J., Snell, A. F., & Foust, S. M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of un-modeled secondary constructs. *Organizational Research Methods*, *2*(3), 233–256.

Higgins, R. L., Snyder, C. R., & Berglas, S. (1990). Self-handicapping: The paradox that isn't. In A. S. R Manstead & M. Hewstone (Ed.), *The Blackwell encyclopedia of social psychology* (p. 515). New York: Plenum. Amazon.com.

Higher Education Research Institute (2002). Cooperative institutional Research Program (CIRP). Higher Education Research Institute, UCLA.

Honora, D., & Rolle, A. (2002). A discussion of incongruence between optimism and academic-performance and its influence on school violence. *Journal of School Violence*, *1*(1), The Haworth Press, Inc.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structure equation modeling: Concepts, issues and applications* (pp. 1–15). Thousand Oaks, CA: Sage Publications. Inc.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks: Sage.

Hu, L. T., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Huffman, K. (2006). *Living psychology* (pp. 340–345). Hoboken: Wiley.

Ibrahim, M. B. (2011). *Lecture notes on statistics*. International Islamic University Malaysia Gombak KL, Malaysia.

Jones, S. L., & Butman, R. E. (1991). *Modern psychotherapies: A comprehensive Christian appraisal* (p. 145). Downers Grove, IL: Intervarsity Press. Jöreskog, K., & Sörbom, D. (1981). Introduction. In A. Kenneth, J. Bollen, & S. Long (1993) (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage.

Jupp, V. (2006). Structured observation. In *Sage dictionary of social research methods*. ISBN: 970857020116.

Kline, R. B. (1998). *Principle and practice of structural equation modeling*. New York: The Guilford Press.

Kline, R.B. (2005) Principle and Practice of Structural Equation Modeling (2nd ed.). New York: The Guilford Press.

Kline, R. B. (2011). *Principle and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Kolditz, T. A., & Arkin, R. M. (1982). An impression management interpretation of the self-handicapping strategy. *Journal of Personality and Social Psychology, 43*(3), 492–502.

Larsen, R. J., & Buss, D. V. (2008). *Personality psychology: Domains of knowledge about human nature* (3rd ed., p. 109). New York: McGraw-Hill.

Lotkowski, V. A., Robbins, S. B., & Noeth, R. J. (2004). *The role of academic and non- academic factors in improving college retention: Act policy report*. Lowa City: ACT Inc.

Lucus, J. W., & Lovaglia, M. J. (2005). Self-handicapping: Gender, race, and status. *Current Research in Social Psychology, 10*(15), 234–249.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Markus, H., & Nurius, P. (1986). Possible selves. *The American Psychologist, 41*(9), 954–969.

Midgley, C., & Urdan, T. (2001). Academic self-handicapping and performance goals: A further examination. *Contemporary Educational Psychology, 26*, 61–75.

Muhammad Taqî-ud-Dîn al-Hîlalî & Muhammad MuhsinKhân. (1426 A.H.) Translation of the Noble Qur'an in the English language. King Fahd complex for the printing of the Holy Qur'ân. Madinah, K.S.A.

National Center for School Engagement (2006). *Quantifying school engagement: Research report*. Promoting attendance, attachment and achievement, Colorado Foundations for Families and Children, USA.

Newman, F. M. (1992). *Student engagement and achievement in American schools*. New York: Teachers College Press. www.getcited.org/pub/102987053.

Newman, L., & Wadas, R. (1997). When stakes are higher: Self-esteem instability ans self-handicapping. *Journal of Social Behavior and Personality, 12*, 217–232.

Nordin, M. S. (2011 & 2012). *Lecture notes on structural equation modeling*. International Islamic University Malaysia, Gombak KL, Malaysia.

Norem, J., & Canter, N. (1986). Anticipatory and post hoc cushioning strategies: Optimism and defensive pessimism in "risky" situations. *Cognitive Therapy and Research, 10*, 347–362.

Nurius, P. S., & Cormier, S. (2003). *Interviewing and change strategies for helpers: Fundamental skills and cognitive behavioral interventions* (5th ed., p. 255). Australia: Thomson Brooks/Cole.

Nurmi, J. (1991). How do adolescents see their future? A review of the development of future orientation and planning. *Developmental Review, 11*, 1–59.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS* (pp. 142–159). Australia: Allen &Unwin.

Pederson, P. B. (1997). *Culture centered counseling: Striving for accuracy*. Thousand Oaks: Sage Publications.

Pederson, P. B. (2007). Ethics, competence and professional issues in cross-cultural counseling 01-Pederson.qxd.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*, 153. In Improving public health prevention with behavioral cognitive science. Report handed to Nathaniel Kosciusko-Morizet, Secretary of State for Strategic Planning and the Development of the Digital Economy. Center for Strategic Analysis 18, rue de Martignac – 757000 Paris cedex 07. www.strategie.gouv.fr

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. ThousandOaks: Sage.

Podolny, J. (1993). A status base model of market competition. *The American Journal of Sociology, 98*(4), 829–872.

Poldony, J. M., & Baron, J. N. (1997). Social networks and mobility. *American Sociological Review, 62*, 673–693. In Kim, H. H. (2001). Social capital, embedded status, and the endorsement effect.

Ravid, R. (2000). *Practical statistics for educators* (2nd ed.). New York: University Press of America, Inc.

Rhodewalt, F. (1990). Self-handicappers: Individual differences in the preference for anticipatory, self-protective acts. In H. Raymond, C. R. Snyder, & B. Steven (Eds.), *Self-handicapping: The paradox that isn't* (pp. 69–106). New York: Plenum.

Rhodewalt, F. (1994). Conceptions of ability, achievement goals, and individual differences in self-handicapping behavior: On the application of implicit theories. *Journal of Personality, 62*, 67–85.

Rhodewalt, F., & Davison, J. (1986). Self-handicapping and subsequent performance: Role of outcome valance and attribution certainty. *Basic and Applied Social Psychology, 7*, 307–322.

Rigdon, E. E. (1998). Structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 251–294). Mahwah: Lawrence Erlbaum Associates, Publishers.

Ritchie, J., & Spencer, L. (1994). *Qualitative data analysis for applied policy research. In analyzing qualitative data*. London: Routledge.

Roebken, H. (2007). The influence of goal orientation on student satisfaction, academic engagement and achievement. *Journal of Research in Educational Psychology, N. 13* 5(3), 679–704.

Ross, S. R., Canada, K. E., & Rausch, M. K. (2002). Selfhandicapping and the five factor model of personality: Mediation between neuroticism and conscientiousness. *Personality and Individual Difference, 32*(7), 1173–1184. doi:10 1016/S011-8869(01) 00079–4.

Saunders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (Research progress report). University of Tennessee Value Added Assessment Center, Knoxville, Tennessee.

Seligman, M. E. P., & Maier, S. F. (1967). Failure to escape traumatic shock. *Journal of Experimental Psychology, 74*, 1–9.

Seligman, M. E. P. (1998). The effectiveness of therapy. *APA Monitor*, *29*, 5. Retrieved December 13, 2004, from www.apa.org/monitor/may98/pc.html.

Showers, C. (1992). The motivation and emotional consequences of considering positive or negative possibilities for an upcoming event. *Journal of Personality and Social Psychology, 63*, 474–483 (Scheier and Carver, 1985).

Sinclair, M. (2007). Editorial: A guide to understanding theoretical and conceptual frameworks. *Evidenced Based Midewifery, 5*(2), 39.

Smith, L., Sinclair, K. E., & Chapman, E. S. (2002). Students' goals, self-efficacy, self-handicapping, and negative affective responses: An Australian Senior School Student Study. *Contemporary Educational Psychology, 27*, 471–485.

Snyder, C. R. (1990). Self-handicapping process and sequelae: On the taking of a psychological dive. In R. L. Higgins, C. R. Snyder, & S. Berglas (Eds.), *Self- handicapping: The paradox that isn't* (pp. 107–150). New York: Plenum Press.

Straus, M. A. (1980). Measuring intrafamily conflict and violence: The conflict tactics. *Journal of Marriage and Family, 41*, 75–88.

Study Circle Guidelines, Co-Curricular Activity Center, Student Development Division. International Islamic University Malaysia, KL, Malaysia.

Sue, D. W., & Sue, D. (2003). *Counseling the culturally diverse: Theory and practice* (4th ed.). New York: John Wiley and sons.

Tanaka, J. S., & Huba, G. J. (1989). A general coefficient of determinant for covariance structure models under arbitrary GLS estimation. *The British Journal of Mathematical and Statistical Psychology, 42*, 233–239.

Wolfe, J. L. (2007). Rational emotive behavior therapy (REBT). In Sharf, R. S. (2011). *Theories of psychotherapy and counseling concepts and cases* (p. 358). Books on Google Play, GANGAGE Learning, GengageBrain.com.

Woods, G. S. (1998). *A study of self-concept as it relates to academic achievement and gender in third grade students.* A Thesis Submitted in partial fulfillment of the requirements of the Master of Arts Degree in the Graduate Division of Rowan University, New Jersey, USA.

Yen, D. H. (1998). *Learned helplessness*. University of Pennsylvania Learned Helplessness Homepage Course. yen@noogenessis.com.

Yick, A. G. (2000). Predictors of physical spousal/intimate violence in Chinese American families. *Journal of Family Violence, 15*(3), 249–267.

Yu, C. L. M., Fielding, R., Chan, C. L. W., Tse, V. K. C., Choi, P. H. K., Lau, W. H., Choy, D. T. K., O, S. K., Lee, A. W. M., & Sham, J. S. T. (2000). Measuring quality of life of Chinese cancer patients. A validation of the Chinese version of the functional assessment of cancer. *Cancer, 88*(7), 1715–1727.

Zuckerman, M., Kieffer, S. C., & Knee, C. R. (1998). Consequences of self-handicapping: Effects on coping, academic performance, and adjustment. *Journal of Personality and Social Psychology, 74*, 1619–1628.

Chapter 4
Implementing Formative Assessment in the Translation Course for English Majors—Taking Beijing Sport University as an Example

Siqi Lv

Abstract Educational assessment as well as studies on it has been playing a pivotal role in language teaching since the concept of "educational assessment" was put forward by R.W. Tyler in the 1930s. As more and more emphasis is put on the development of students' ability and the process of teaching, formative assessment, which differs from the traditional result-oriented summative assessment, is being widely applied. Although the benefits brought by and importance of formative assessment have been confirmed, most of the researches are carried out only in college English classes, with little practice in English majors. This thesis intends to implement formative assessment in the translation course for English majors, aiming at finding out its impacts on students' translation ability as well as learning capabilities, and further providing suggestions on future studies in this area. The research finds out that the students' role has changed from a listener to a participant, while the teacher from an authority to a guide. The students' translation ability has been improved after the research and they have made progress in other learning capabilities like self-reflection, critical thinking and learning autonomy, which are crucial for future studies in other areas.

Keywords Formative assessment • Translation course • Translation ability • Learning capability

4.1 Introduction

Educational assessment, which mainly includes summative and formative assessment, has been playing a pivotal role in supervising the teaching process, checking the teaching effects and improving communication between students and the teacher

S. Lv (✉)
School of Foreign Languages, Beihang University, Foreign Languages Department, Beijing Sport University, Beijing, China
e-mail: bsu_kenny@foxmail.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_4,

(Yang 2004). A well-confined evaluating system is helpful to teachers, students as well as administrators (Wen 2007). The traditional summative assessment, which focuses on the final scores or the learning results, has been applied in China for a long time (Xie 2008). Not until in recent years have researches concerned about formative assessment (ibid.). *Curriculum Requirements for College English Teaching* (*Requirements* hereafter) published in 2004 ushered a new era for assessment at college level. It emphasizes the importance of formative assessment and requires an evaluating system which contains both formative and summative assessment (Luo 2010). The former assessment system which focuses on the products should be applied in combination with a process-oriented one.

Presently, formative assessment is more and more widely accepted throughout China. However, most of the researches are still carried out in college English classes, with little practice concerning about English majors. Teachers who have not fully understood the process of formative assessment may be reluctant to apply it in their teaching as well. Former studies generally focused only on the immediate effects of formative assessment instead of students' improvement on their autonomous learning as well as the self-reflective ability, which are of great value in the modern time. Therefore, this article intends to apply formative assessment to the translation course of junior English majors at Beijing Sport University (BSU), expecting to find out whether the students' translation ability as well as learning capability can be improved or not. Hopefully, this study could provide some feasible suggestions on the teaching of English major's translation course and the application of formative assessment to further studies.

The remainder of this chapter is organized as follows. Section 4.2 is the literature review of the types and development of educational assessment, after which formative assessment in particular will be described. In Sect. 4.3, research design and methodology will be listed, which explains how the experiment is going to be carried out. Section 4.4 deals with the quantitative and qualitative results. And in the end, Sect. 4.5 gives a discussion, indicating the major findings, limitations and suggestions for further studies as well.

4.2 Formative Assessment

Long categorizes the assessment into four types—summative assessment, product assessment, formative assessment and process assessment (Long 1984). According to Luo, general types of assessment include diagnostic assessment, formative assessment, summative assessment, sizing up assessment, instructional assessment and official assessment (Luo 2003). Summative assessment has been traditionally used to determine whether a program is successful or not (no matter what the definition of "success" is), but it cannot show how and why it is successful or provide guidance for further teaching; nevertheless, formative assessment is always used as a means to achieve success (Cao et al. 2004).

4.2.1 Studies on Formative Assessment Abroad and in China

Scriven first puts forward the concepts of formative assessment and summative assessment (Chen 2008). Bachman points out that the differences between formative assessment and summative assessment lie in the time when the assessment is implemented and what the purposes of the implementation are (Xie 2008). As the focus of education transformed from the teaching results to students' learning process after the educational reform in 1990s, new modes on assessment come out one after another (ibid.). Herbert offers a practical approach to using the portfolio in assessment of elementary level (Chen 2008). Gimenez puts forward a formative assessment mode which is applicable in ESP, and testifies the importance of formative assessment as well as its functions to EFL teaching (Xie 2008). Lynch, Genesee and Upshur provide a framework as well as a practical guide for language evaluation in their books respectively.

Testing has had a long history in China, but it is not until recently that the topic of assessment arouses public attention. Hu first explains in detail the purposes and principles of formative assessment, and discusses its application to college English teaching (ibid.). Requirements calls for a combination of formative and summative assessment and thereafter, more and more language educators and scholars begin to introduce and practice formative assessment. Cao et al. apply formative assessment to the English writing course and find out it is helpful to enhance students' learning autonomy (Cao et al. 2004). After that, formative assessment is implemented in the Internet-based college English courses by Zhou and Qin (Xie 2008). Luo and Peng introduce the method of portfolio into formative assessment, pointing out the improvement of students' ability of self-reflection (ibid.).

4.2.2 Definition, Types and Tools of Formative Assessment

The terms "formative assessment" and "summative assessment" are first introduced by Scriven who has claimed the difference between them lies in whether the focus is the process or the product (Scriven 1967). According to Black and William, "assessment refers to all those activities undertaken by teachers, and by the students in assessing feedback to modify themselves. Such assessment becomes 'formative assessment', when the evidence is actually used to adapt the teaching to meet their goals" (Black and William 1998). Garrison and Ehringhaus point out that formative assessment is part of the instructional process which informs both teachers and students of student understanding at a point when timely adjustments can be made (Garrison and Ehringhaus 2007).

There are roughly three types of formative assessment: self-assessment, peer-assessment and teacher assessment. First, self-assessment, according to Harris, is "a self-directed and self-determining learner setting his or her own assessment criteria, judging his or her learning process (or products) against these criteria, and making

decisions based on these judgments" (Harris and Chris 1986). Students tend to be more motivated and have a better understanding of them by self-evaluation. Second, peer-assessment is considered by Bound as an activity usually combined with self-assessment and "in the right circumstances call considerably self-assessment" (Bound 1995). Knowing about others' judgment is of much help to recognize one's own weakness, so that proper adjustments could be made in the process of learning. Third, evaluation given by the teacher is absolutely crucial to students. As an authority in Chinese classes, a teacher should be aware that their feedback plays a pivotal role in students' learning. If a teacher encourages a lot and more importantly, points out students' strengths and weaknesses, students will get the idea that they are cared about and thus become more confident and find ways to improve the performance more efficiently (Li 2010).

Tools used in the process of formative assessment always involve tests, journals, observation, questionnaires and interviews. A test is a task or set of tasks that elicits observable behavior from test takers and yields scores that represent attributes or characteristics of individuals (Genesee and Upshur 1996). In formative assessment, a pre- as well as a post- test is usually employed to see whether students' academic performance has been enhanced. In addition, journals are writing records of students' learning. Regular journals provide teachers with opportunities to assess their students' ability to express themselves personally in writing using the second language without the pressure that students may feel during class activities (ibid.). Furthermore, researchers have devised sophisticated formal methods of observing in order to describe and understand EFL teaching better, but such methods usually take too much time and thus are not used by teachers in the class of formative assessment. Instead, an informal and lasting observation could be applied by teachers so that they can assess what students have and have not learned, the effectiveness of particular teaching strategies and so on (ibid.). Finally, questionnaires and interviews are structured and formal assessing tools. They are used after instruction to gather information about the effectiveness of a unit or an entire course—information such as students' general impressions on the course and its various components and their satisfaction with their achievements (ibid.).

4.3 Methodology

4.3.1 Participants

Forty junior English majors were selected for the experiment with a purpose to examine whether formative assessment could assert some influence on their translation as well as learning abilities. They were taught by the same teacher and the whole research lasted 8 weeks, namely, the second half of the semester, with 16 teaching hours in total.

4.3.2 Methods

Self-assessment and peer-assessment: From Week 8 on, students were required to finish their handouts sent by the teacher before each class, with their self-evaluation as well as one or two classmates' evaluation written behind. They were encouraged to rethink about and summarize their own weakness and progress in self-assessing, and to give critical reflections on others' translation during the peer review. Guidance from the teacher was provided when necessary.

Teacher assessment: In this research, the teacher gave her evaluation in two ways: during the classroom teaching and in students' handouts. More emphasis was put on the feedback and assessment given to the students in class. Group and class discussions were held regularly, through which the students were able to have a deep understanding and gain enough feedback. After reviewing their assessment in the handouts, the teacher would write down her evaluation and give it back in person or in class.

4.3.3 Procedures

Achievement tests: Two achievement tests were given to the students in W8 and W16 respectively. Both the pre- and post- tests were E-C translation passages of the same degree of difficulty selected from the previous TEM-8 translation part. Scores were given by the author according to the scoring standards of TEM-8, after which quantitative data would be concluded to see whether students' translation ability has improved or not during the past 8 weeks.

Journals: Students were encouraged to write a journal every 2 weeks on the following questions:

1. What kind of class activity did I participate in? How did I think of it?
2. What have I learned during the past 2 weeks?
3. Was there any change in my learning strategies recently?
4. Are the peer-assessment and self-assessment helpful to improve my translation skills? What can I learn from my classmates' evaluation?

If there were no reflection concerning the questions above, they could write anything that they wanted to share with the teacher.

Classroom observation: Observations were made by the teacher on students' classroom participation, translation quality, ability of thinking critically and the general class atmosphere. After each class, the teacher's observing results and her perception of changes would be given to the author and then be recorded.

Questionnaires: Questionnaires on students' opinions towards formative assessment were used at the end of the research. The five-scale questionnaire was designed in Chinese by the author herself, containing 15 multiple choice questions and two subjective questions. The framework of the questionnaire is shown in Table 4.1.

Table 4.1 Framework of the questionnaire

Aspect	Item
General impression	1, 2, 3, 4, 5, 6
Changes in ability	7, 8, 9, 10, 11
Feedback and suggestions	12, 13, 14, 15

In-depth interviews: Six students were chosen for structured interviews at the end of the research. Questions on their opinions of formative assessment, achievements after the practice and suggestions on future teaching were asked in the interviews.

4.3.4 Data Collection

The data collected included quantitative statistics and qualitative description.

Quantitative data were concluded from the results of the two achievement tests as well as the questionnaires. Microsoft Excel 2010 and IBM SPSS 19.0 were used in collecting and calculating the scores.

Qualitative description was summarized from students' assessment, journals and interviews, and the teacher's reflection via her classroom observation.

4.4 Results

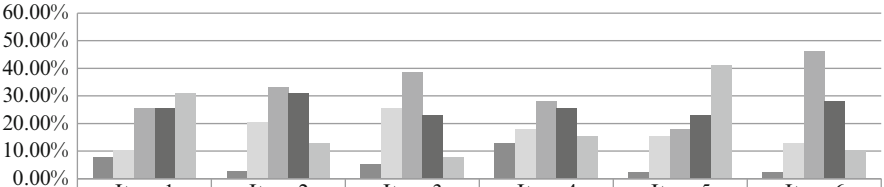
4.4.1 Quantitative Data

4.4.1.1 Results from Questionnaires

Forty questionnaires were given out to the students in total and 39 were effectively retrieved. The statistics were analyzed in accordance with the framework of the questionnaire.

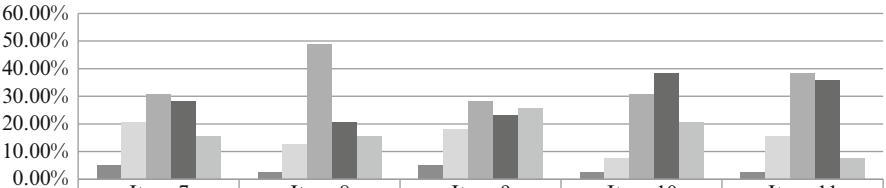
Students' general impression on formative assessment: Generally speaking, most of the students were interested in and had learnt a lot from the process of formative assessment, as shown in Item 1 and Item 2. Although according to Item 3, there were still more than 30 % students who were totally or always not involved in it, 41 % students totally agreed in Item 5 that evaluating others' work was helpful to improve their own translation ability. In Item 4, there were 40 % of the students holding positive opinions towards teacher assessment, which was greater than those who didn't. For Item 6, 84.7 % students in total confessed it was at least sometimes motivating to see the assessment given by others (Fig. 4.1).

The improvement of students' translation ability: In Item 7 which dealt with learning autonomy, 43.6 % of the students agreed they became more active in learning while only 5.1 % totally disagreed and 20.5 % always disagreed. Almost half of the students thought they could sometimes provide accurate and objective



	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	7.70%	2.60%	5.10%	12.80%	2.50%	2.50%
2	10.30%	20.50%	25.60%	18.00%	15.40%	12.80%
3	25.60%	33.30%	38.50%	28.20%	18.00%	46.20%
4	25.60%	30.80%	23.10%	25.60%	23.10%	28.20%
5	30.80%	12.80%	7.70%	15.40%	41.00%	10.30%

Fig. 4.1 Statistics of Item 1–6. *Note: 1* totally disagree, *2* always disagree, *3* sometimes agree, *4* always agree, *5* totally agree

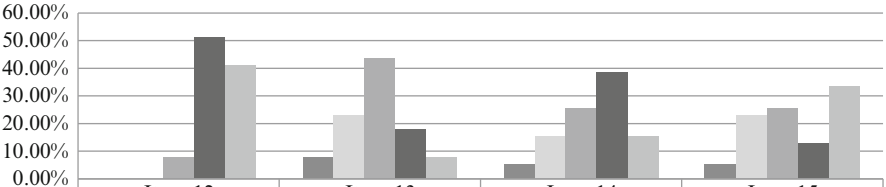


	Item 7	Item 8	Item 9	Item 10	Item 11
1	5.10%	2.60%	5.10%	2.50%	2.50%
2	20.50%	12.80%	18%	7.70%	15.40%
3	30.80%	48.70%	28.20%	30.80%	38.50%
4	28.20%	20.50%	23.10%	38.50%	35.90%
5	15.40%	15.40%	25.60%	20.50%	7.70%

Fig. 4.2 Statistics of Item 7–11. *Note: 1* totally disagree, *2* always disagree, *3* sometimes agree, *4* always agree, *5* totally agree

judgments on others' translation in Item 8, with only 2.6 % of them did not think so. No apparent differences were shown in Item 9, but it was safe to say that those who held positive views were more than those with negative opinions. In Item 10, only one tenth of the students totally or always disagreed that self-assessment could help find out their own weaknesses, while 38.5 % students always and 20.5 % totally thought positively. Item 11 demonstrated that in general, more students would like to reflect on their study and make summaries regularly (Fig. 4.2).

Feedback and suggestions: It can be easily concluded from Item 12 that regular self-assessment or self-reflection was of significance to all the students. However, Item 13 showed that seven tenths of students held that practice was as important as, if not more than, the process of assessing. Thus in Item 14, still more than half of the students thought their translation ability would remain the same without the assessing practice. As to whether it was necessary to apply formative assessment to future teaching, only one third totally agreed, with 23.1 % always disagreed and 25.6 % sometimes agreed (Fig. 4.3).



	Item 12	Item 13	Item 14	Item 15
1	0	7.70%	5.10%	5.20%
2	0	23.10%	15.40%	23.10%
3	7.70%	43.60%	25.60%	25.60%
4	51.30%	17.90%	38.50%	12.80%
5	41%	7.70%	15.40%	33.30%

Fig. 4.3 Statistics of Item 12–15. *Note*: *1* totally disagree, *2* always disagree, *3* sometimes agree, *4* always agree, *5* totally agree

4.4.1.2 Results from Two Achievement Tests

According to the scoring system of TEM-8, there are five rates of scores: (1) 10–9: excellent translation; (2) 8–7: good translation with few inaccuracies; (3) 6–5: passable translation with some inaccuracies; (4) 4–3: inadequate translation with frequent inaccuracies; and (5) 2–1: poor translation. Both the pre- and post- tests were scored by the author, after which a comparison was made as follows:

As shown in Fig. 4.4, there was no student getting four points in the post-test, while more people in the post-test got five and nine than in the pre-test. As the percentage of six points decreased, the proportion of seven and eight points increased significantly. Thus it is safe to conclude that after the practice of formative assessment in the latter half semester, students' general level has been improved.

In addition, a Paired Sample T-test was made between the pre-test and post-test. The statistics is shown in Table 4.2:

According to the result above, the mean was −0.5 with a standard deviation of 0.507. The lower and upper intervals of the difference were −0.667 and −0.333 respectively, which contained no 0 and meant there was a significant difference between the two variables. The 2-detailed significance was 0, which was much less than 0.05, meaning there was a significant difference between the pre-test and post-test as well.

4.4.2 Qualitative Data

4.4.2.1 Students' Feedback from Assessments, Journals and Interviews

The self-assessment and peer-assessment in students' handouts could directly reflect their learning progress and to what extent they had mastered the translation

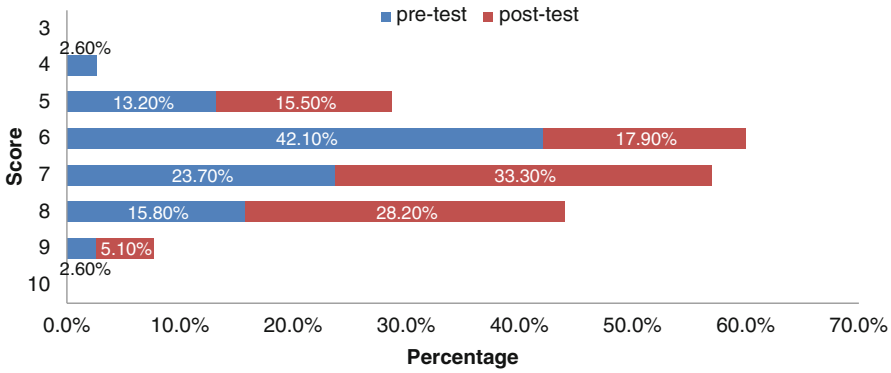


Fig. 4.4 An analysis of the test scores

Table 4.2 Result of the paired samples test

		Paired differences					t	df	Sig. (2-tailed)
					95 % confidence interval of the difference				
		Mean	S.D.	S.E. Mean	Lower	Upper			
Pair 1	Pre-test – Post-test	−.500	.507	.082	−.667	−.333	−6.083	37	.000

Note: S.D. standard deviation, *S.E. Mean* standard error mean. If 0 is not included in the 95 % confidence interval of the difference, or the 2-tailed significance is less than 0.05, it can be concluded that there are significant differences between the two variables (Qin 2003)

skills. Many of them thought the assessment process was "useful and enjoyable", through which they had learnt to compare different translation versions and find out their particular shortcomings. They were able to appreciate others' translation, make reasonable judgments and "learn from them".

In the journals, some students pointed out that "it's better for both the teacher and students to participate in the interaction more" and that "more exercise should be given after class rather than before class". Such kinds of suggestions were quite helpful for the teacher to know about the students' reflections and adjust her teaching in time. In addition, most of the students wrote out their learning progress like "I really learnt a lot about the weakness that I've not noticed", "apparently my translation skills has improved" and "the sense of translation and ability improved apparently" etc.

Interviews were used as a supplementary tool of questionnaires. Students' answers to the short questions in questionnaires and interview questions were summarized as follows. First, views towards self-assessment, peer-assessment and teacher assessment varied a lot among students. Some people thought that it was useless to do self-assessment, while reviewing others' work and knowing classmates' evaluation on them were of great value. In terms of teacher assessment, most of the students considered it was not enough. Moreover, students' translation

ability has been improved in general since many of them confessed that they have learnt to reconsider and summarize their studies routinely. Some suggestions on further teaching of the translation course were also provided. Handouts were strongly expected to contain both the new translation skills to be taught in the next class and the one explained in the previous class. Also, not a few students hoped the teacher could communicate more with them in person and provide advices concerning the future at the same time.

4.4.2.2 Teacher's Feedback via Classroom Observation

The teacher has felt a significant change in the classroom atmosphere between the two half semesters. First, students participated more actively in class discussion, which however rarely happened in the former half semester. In the first few weeks of the semester, the teacher felt bored and no passion because of classroom silence; but after W8, heated discussions in class became essential. Second, students' ability of critical thinking has improved. Faced with two to three different versions, they were able to analyze and judge which one was better translated. This kind of ability is crucial not only in translation learning, but in other courses. Third, it was easier for the teacher to find out students' shortcomings on the basis of their performance, which led to timely suggestions for them. If students did not say anything in or after class, the teacher would never know to what extent they have learnt or whether it was necessary to adjust her teaching arrangements.

Yet, regardless of the positive changes mentioned above, several problems should not be neglected. For example, the time spent on translation was far less than enough for some students, which could be reflected in the quality of their handouts. In addition, not all students persisted in the practice of assessing as well as journal writing. Careless and hasty assessments appeared later in the semester. Lastly, extra-curriculum exercises were still needed so that the skills taught in class could be practiced in time.

4.5 Discussion

This current study is carried out in the combination of both a quantitative and a qualitative approach. The previous results could, as expected, shed light on future language teaching as well as learning. In this part, the article will be concluded by a brief discussion of the data given above, followed by a conclusion and some suggestions for further study.

Quantitative statistics can reflect the facts most directly and objectively. In quite a few items of the questionnaire, there are significantly more people who tend to confess the positive effects brought by formative assessment. Although we cannot see significant differences in a few items, or even there are more people hold negative opinions, it shouldn't be ignored that there do exist at least some students taking positive answers. The reason why formative assessment is not favored by all might be some misunderstandings during the implementing process carried out with

the help of the teacher. Also, students' attendance of the class may affect their learning as well as the comprehension of formative assessment. The test scores, which accord with the normal distribution, is an additional proof of the effectiveness of formative assessment. However, there might be doubts that the improvement of translation skills is due to students' natural learning instead of the assessing process. This problem, as has been predicated, is solved by a qualitative study through in-depth interviews. The qualitative data will be discussed in the following.

In terms of the qualitative approach, self- and peer- assessments as well as journals are required with the expectation of letting students reconsider the previous learning as well as make a regular self-reflection. Usually, students tend to express freely in their writings instead of face-to-face dialogues. Most students participate actively in the assessing process, despite some of them become careless later, as the teacher observes. It is also worth pointing out that many students become much more confident when answering questions in class, according to their own account. In the interviews, the students not only talk about their improvements, but also give suggestions on further teaching—for example, more specific and frequent feedback from the teacher is strongly expected. Besides, the teacher's feedback plays an important role since it is she who implements the whole process and who is the direct observer of all the proceedings. In spite of communicating and understanding difficulties, the teacher's work prove to be acknowledgeable and the results convincible.

In short, we can finally come to the following conclusions. First, both the teacher's and students' roles in class have changed. The teacher has transformed from a controller to a guide with the implementation of formative assessment, which provides us a brand-new idea that students are part of the assessing program. Most of the students hold a positive attitude towards formative assessment and readily participate in it. They are no longer passive listeners, but active learners. Then, there are significant improvements concerning the students' translation abilities. The classroom is no more silent as students are more confident to share their versions or even argue with the teacher for a better translation. Also they have made learning achievements as shown in the comparison between the pre- and post-tests. Finally, major progress in students' learning capabilities, including learning autonomy, self-reflection and critical thinking is gained. The students become more actively involved in their learning, know how to reconsider about themselves regularly and think critically. All of these abilities are pivotal not only in translation learning, but also other courses.

Although great efforts have been put into the research, there are still inevitable limitations. More studies with a longer time span and a larger sample size are expected, and other tools like conferences and portfolios could be used in future research as well. It is also worth trying to attest how effective it is to combine summative and formative assessment in a wide range of courses. Hopefully this empirical research could provide some clues and feasible suggestions for further foreign language teaching as well as researching.

Appendix A. Questionnaire on Students' Opinions Toward Formative Assessment in Translation Course

Note. In order to make it easily understood by Chinese students, the questionnaire is designed in Chinese and all the items are translated as follows.

Item 1: I prefer the translation course of the latter half semester.
Item 2: I have learnt a lot from the self- and peer- assessment.
Item 3: I participate in the process of self- and peer- assessment actively.
Item 4: I think teacher's assessment given to us is very helpful.
Item 5: Evaluating others' work helps me to figure out my weaknesses and improve my skills.
Item 6: Watching other's assessment on me is very motivating.
Item 7: I become more interested in translation and know how to learn actively.
Item 8: My ability of critical thinking has improved and I am able to evaluate others' work objectively and accurately.
Item 9: I become more active in the class discussion because of the pre-class assessment.
Item 10: Through self-assessment, I am able to find out my progress and weakness, and ways to improve.
Item 11: I spend more time in self-reflection, rethinking and summarizing my learning routinely.
Item 12: I think regular self-reflections are helpful to my study.
Item 13: I can learn translation effectively with only large amounts of practice and no assessing.
Item 14: My translation ability would be the same as what it is now if there was no process of assessing in the latter half semester.
Item 15: I suggest formative assessment be applied in next semester.

Appendix B. Interview Questions

Q1. What is the difference between the translation course before and after Week 8?
Q2. What can you learn from self-assessment, peer-assessment and teacher's assessment? Please explain it respectively.
Q3. Will you reconsider your strengths and weaknesses in translation after each class? What benefits have you got from it? If not, tell the reason.
Q4. Have you made any progress in your translation studies? Any change in motivation? What should be done to improve the teaching of translation?

Appendix C. Pre-test (TEM-8 2002)

Part V Translation

Section B English to Chinese

Translate the underlined part of the following text into Chinese.
The word "winner" and "loser" have many meanings. When we refer to a person as a winner, we do not mean one who makes someone else lose. To us, a winner is one who responds authentically by being credible, trustworthy, responsive, and genuine, both as an individual and as a member of a society.

Winners do not dedicate their lives to a concept of what they imagine they should be; rather, they are themselves and as such do not use their energy putting on a performance, maintaining pretence, and manipulating others. They are aware that there is a difference between being loving and acting loving, between being stupid and acting stupid, between being knowledgeable and acting knowledgeable. Winners do not need to hide behind a mask.

Winners are not afraid to do their own thinking and to use their own knowledge. They can separate facts from opinions and don't pretend to have all the answers. They listen to others, evaluate what they say, but come to their own conclusions. Although winners can adore and respect other people, they are not totally defined, demolished, bound, or awed by them.

Winners do not play "helpless", nor do they play the blaming game. Instead, they assume responsibility for their own lives.

Appendix D. Post-test (TEM-8 1999)

Part V Translation

Section B English to Chinese

Translate the underlined part of the following text into Chinese.
In some societies people want children for what might be called familial reasons: to extend the family line or the family name, to propitiate the ancestors; to enable the proper functioning of religious rituals involving the family. Such reasons may seem thin in the modern, secularized society but they have been and are powerful indeed in other places.

In addition, one class of family reasons shares a border with the following category, namely, having children in order to maintain or improve a marriage: to hold the husband or occupy the wife; to repair or rejuvenate the marriage; to increase the number of children on the assumption that family happiness lies that

way. The point is underlined by its converse: in some societies the failure to bear children (or males) is a threat to the marriage and a ready cause for divorce.

Beyond all that is the profound significance of children to the very institution of the family itself. To many people, husband and wife alone do not seem a proper family—they need children to enrich the circle, to validate its family character, to gather the redemptive influence of offspring.

Children need the family, but the family seems also to need children, as the social institution uniquely available, at least in principle, for security, comfort, assurance, and direction in a changing, often hostile, world. To most people, such a home base, in the literal sense, needs more than one person for sustenance and in generational extension.

References

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

Bound, D. (1995). *Enhancing learning through self-assessment*. London: RoutledgeFalmer.

Cao, R., Zhang, W., & Zhou, Y. (2004). Implementation of formative evaluation in an EFL writing course for Chinese college non-language majors. *Foreign Language Education, 5*, 82–87.

Chen, J. (2008). *An empirical study on application of formative assessment to English majors in China*. Unpublished master's dissertation, Nanchang University, Nanchang.

Garrison, C., & Ehringhaus, M. (2007). *Formative and summative assessment in the classroom*. Retrieved from: http://www.amle.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Harris, D., & Chris, B. (1986). *Evaluating and assessing for learning*. New York: Kogan Page Limited.

Li, H. (2010). *On cultivating students' autonomous learning ability and application formative assessment into college English teaching*. Unpublished master's dissertation, Shanghai International Studies University, Shanghai.

Long, M. H. (1984). Process and product in ESL program evaluation. *TESOL Quarterly, 18*, 409–425.

Luo, S. (2003). *Formative assessment in English classroom teaching*. Beijing: Foreign Language Teaching and Research Press.

Luo, Q. (2010). Formative assessment approach and mechanism in network—based college English teaching. *Journal of Sichuan College of Education, 26*(3), 95–97.

Qin, X. (2003). *Quantitative studies in the research of foreign language teaching*. Wuhan: The Huazhong University of Science and Technology Press.

Scriven, M. (1967). *The methodology of evaluation*. West Lafayette: Purdue University.

Wen, Q. (2007). *A study on the implementation of formative assessment in EFL teaching for English majors*. Unpublished master's dissertation, Shanghai International Studies University, Shanghai.

Xie, F. (2008). A review of formative assessment and its application in foreign language teaching. *The Border Economy and Culture, 6*, 119–120.

Yang, W. (2004). Views on application of formative classroom assessment. *Journal of Guangzhou University (Social Science Edition), 3*(12), 83–86.

Chapter 5
Further Implementation of User Defined Fit Statistics

Daniel Urbach

Abstract This paper further investigates results from Adams and Wu (J Appl Meas, 10(4), 355–370, 2009) who developed a User Defined Fit Statistic to test the fit of linear combinations of items (or parameters) using linear combinations of individual contributions to the sufficient statistics and comparing them to their expectations. The User Defined Fit Statistic has the ability to identify violations of local independence and/or violations of uni dimensionality using a priori knowledge of the data. In a lot of cases, individual item (or parameter) residual based fit statistics are unable to identify such model violations. Simulations showed the User Defined Fit Statics' ability to identify item response model violations, when local item dependence and multi dimensionality was present in simulated data. Two real datasets were also analysed, where local item dependence and multi dimensionality were again identified, displaying how useful these fit statistics can be (with a priori knowledge of the data) in identifying such model violations.

Keywords Item Response Theory • Rasch Measurement, item fit • User Defined Fit, item dimensionality, local item dependence

5.1 Introduction

Individual item fit statistics in Item Response Theory (IRT) models, such as the Wright and Masters (1982) infit mean square and outfit mean square statistics are one of the most commonly used fit statistics in IRT calibration analyses. Such fit statistics help identity individual items that do not fit model expectations. However these fit statistics are also restricted to the item level and can have very limited

D. Urbach (✉)
Psychometrics and Methodology, Australian Council for Educational Research,
19 Prospect Hill Road, Camberwell, VIC 3124, Australia
e-mail: daniel.urbach@acer.edu.au

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_5,

ability in detecting hypothesized model violations consisting of groups of items (such as local item dependence and multi dimensionality). For example, if a 50 item test is made up of 45 items that belong to one dimension and five items that belong to another dimension, then individual item fit statistics may be able to identify the uni dimensionality violation. But if the number of items on each dimension is closer to a 50:50 split then these individual item fit statistics may not be able to pick up the uni dimensionality violations.

Adams and Wu (2009) developed a the more flexible User Defined Fit Statistic to test the fit of linear combinations of items (or parameters in the case of models with facets) using linear combinations of individual contributions to the sufficient statistics and comparing them to their expectations. These User Defined Fit Statistics easily allow for the fit of groups of items (or parameters) to be tested which provides a much more powerful option compared to the individual item (or parameter) fit statistics, when needing to test specific hypotheses. Examples of such groups could be the sub strands of a test, or different item units or item types in a particular test or in a model with raters as facets the groups could be different groups of markers, etc.

The next section discusses the User Defined Fit Statistics and examines in particular, the use of a design matrix that is utilised to construct them in ACER ConQuest (2007). This follows the results of the simulation analyses, one data set where item dependence was induced in the data and one data set where multi dimensionality was induced in the data. Two real datasets are also analysed to which the User Defined Fit Statistics are applied to investigate item dependence for one data set and confirm multi dimensionality in the other data set.

5.2 Generalised Item Response Model and User Defined Fit Statistics

Derived from the Wright and Masters (1982) fit statistics are the fit statistics extended to be used with the generalised item response theory models of Adams, Wilson and Wu (Adams and Wilson (1996), Adams et al. (1997), Wu et al. (2007)). The fit statistics in the generalised item response model are constructed to refer to each "parameter" rather than each "item" to accommodate the fact that the model parameters may not refer to items in all instances (for example, they may be facets). Wu (1997) and Wu et al. (2007), derived these residual based fit tests, which are based on the sufficient statistics for the parameters (which in a simple model with only items are based on each item score).

The generalised item response theory models follow the notion of a design matrix and a score vector that map the item responses to the specified underlying IRT model. The fit statistics of the generalised item response theory models are based on the linear combination of item responses, $\mathbf{A}'_p\mathbf{x_n}$, which provide a fit statistic for each parameter. Where $\mathbf{A}$ is a design matrix of 0's and 1's (and the 1's indicate each parameter p) and $\mathbf{x_n}$ is the response vector of person n. These linear combinations per parameter can be taken advantage of, to produce more generalised fit statistics,

Table 5.1 Design matrix, **A**, for a test with 5 items

	Parameters				
Items and Categories	P1	P2	P3	P4	P5
Item 1 Category 0	0	0	0	0	0
Item 1 Category 1	1	0	0	0	0
Item 2 Category 0	0	0	0	0	0
Item 2 Category 1	0	1	0	0	0
Item 3 Category 0	0	0	0	0	0
Item 3 Category 1	0	0	1	0	0
Item 4 Category 0	0	0	0	0	0
Item 4 Category 1	0	0	0	1	0
Item 5 Category 0	0	0	0	0	0
Item 5 Category 1	0	0	0	0	1

Table 5.2 Design matrix, **F**,for two fit tests

	Fit tests	
Items and Categories	F1	F2
Item 1 Category 0	0	0
Item 1 Category 1	1	0
Item 2 Category 0	0	0
Item 2 Category 1	1	0
Item 3 Category 0	0	0
Item 3 Category 1	1	0
Item 4 Category 0	0	0
Item 4 Category 1	0	1
Item 5 Category 0	0	0
Item 5 Category 1	0	1

where any linear combination of item responses may be used. If $\mathbf{F}_u$ is any vector of the same length as $\mathbf{A}_p$, then $\mathbf{F'}_u\mathbf{x_n}$ is a linear combination of item responses just like $\mathbf{A'}_p\mathbf{x_n}$ is. The expectation and variance of $\mathbf{F'}_u\mathbf{x_n}$ can be computed in the same manner as is for $\mathbf{A'}_p\mathbf{x_n}$, and therefore a fit statistic for each parameter, u can be constructed in exactly the same way as is done for each parameter p.

For example, consider the design matrix when there are five dichotomous items in Table 5.1.

Here, the first column of $\mathbf{A}$ is $\mathbf{A'}_1 = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$ and $\mathbf{A'}_1\mathbf{x_n}$ gives the item response of person n to item 1, or in another words the contribution of person n to the sufficient statistic for the first item parameter. Therefore, $\mathbf{A'}_p\mathbf{x_n}$ gives the contribution of person n to the sufficient statistic for item parameter p. The design matrix and response vector is used as a basis for producing the fit statistics for each parameter, as shown by Adams and Wu (2009).

For a User Defined Fit test, the design vector $\mathbf{A}_p$, can be replaced by any arbitrary vector $\mathbf{F}_u$. Hence, the design matrix $\mathbf{A}$ of p fit tests can be replaced by any design matrix $\mathbf{F}$ of u fits tests. For example, the design matrix shown above can be constructed to give two fit tests. This design matrix $\mathbf{F}$ could then be as follows in Table 5.2.

This provides two fit design vectors; $\mathbf{F'}_1 = (0,1,0,1,0,1,0,0,0,0)$ and $\mathbf{F'}_2 = (0,0,0,0,0,0,0,1,0,1)$, where $\mathbf{F'}_1\mathbf{x_n}$ gives the total score of the first three parameters for person n and $\mathbf{F'}_2\mathbf{x_n}$ gives the total score of the last two parameters for person n. The design matrix and response vector is then used as a basis for producing the User Defined Fit Statistic for each design vector, in the same manner as is done at the parameter level, as shown by Adams and Wu (2009).

This is how the User Defined Fit Statistic can be constructed, and fit tests of item (or parameter) groups can be specified.

In practice these fit tests can be used in different ways. These fit tests could be produced for groups of items that are suspected or expected to belong to different dimensions. If combinations of items are assumed to breach local independence assumptions, a fit test could be constructed containing such item groups (for example sets of items that refer to the same stimulus of text). If a model is estimated with Raters as facets, fit tests containing different Rater groups could also be tested for to test for Rater dependencies.

5.3 Simulations

The capacity of these User Defined Fit Statistics to identify violations of local independence and uni dimensionality compared to the generalised item fit statistics are investigated through various simulations below, using the simple dichotomous Rasch (or one parameter) model.

To say that two items are locally independent intuitively means that the item response of the first item does not increase nor decrease the response probability of the second item and vice versa, after accounting for the latent trait that is being measured. More formally for two items, item $i - 1$ and item i, after accounting for the latent trait, the response probability of both item $i - 1$ and item i is equal to the response probability of item $i - 1$ multiplied by the response probability of item i.

Lord and Novick (1968), explain that under local independence, item scores in social science tests are related to each other for the total group of examinees, but that items scores are related only through the latent variable(s) being tested. Goldstein (1980), states that in the case of mental health tests, independence in the response sets *between* individuals is far more likely to occur than *within* individuals, where the possibility of mutual dependencies exist.

An example of local dependence, between a dichotomous pair of items, is when the response to item i may depend on whether the response to item $i - 1$ is correct or not. If the response to item $i - 1$ is correct, then the probability of getting item i correct could be higher than expected. And vice versa, if the response to item $i - 1$ is incorrect, then the probability of getting item i correct could be lower than expected.

5.3.1 Inducing Local Dependence

A method of inducing local dependence is described by Andrich and Marais (2005). This method is used in this paper to induce local independence violations into simulated data sets to investigate the effect on the User Defined Fit Statistics. This method simply introduces an extra parameter into the Rasch model.

Let:

δ_i = Generating Item Difficulty for Item i

β_n = Generating Person Ability for Person n

X_i = Binary ("0" or "1") response for Item i by a particular person

Without any local item dependence:

$$\Pr(X_i = 1) = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

Introducing Dependence of item i on item $i - 1$:

$$\text{If } X_{i-1} = 1;\ \text{then for this individual } \Pr(X_i = 1) = \frac{e^{(\beta_n - (\delta_i - \rho))}}{1 + e^{(\beta_n - (\delta_i - \rho))}}$$

$$\text{If } X_{i-1} = 0;\ \text{then for this individual } \Pr(X_i = 1) = \frac{e^{(\beta_n - (\delta_i + \rho))}}{1 + e^{(\beta_n - (\delta_i + \rho))}}$$

If the response on item $i - 1$ is correct then the difficulty of item i becomes lower $(\delta_i - \rho)$ and if the response on item $i - 1$ is incorrect then the difficulty of item i becomes higher $(\delta_i + \rho)$.

For $\rho > 0$ logits, there exists a dependence effect

For $\rho = 0$ logits, there exists NO dependence effect

For $\rho < 0$ logits, there exists an opposite dependence effect (i.e. item i is more difficult if item $i - 1$ is correct or item i is easier if item $i - 1$ is incorrect – this case is ignored in the analyses below).

5.3.2 Simulation 1: Testing Violations of Local Independence

Using the method of inducing local item dependence, described by Andrich and Marais (2005), data sets were generated for 1,000 cases with ten dichotomous items, where item 2 has an induced dependence on item 1, item 4 has an induced dependence on item 3, and so on. 1,000 replications were undertaken. Case abilities were generated from a uni variate normal distribution with mean 0 and variance 1. Item difficulties were generated from a uniform distribution with a range of

Table 5.3 Design matrix FIT A – Infit: individual items

	Fit tests									
Items and Categories	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Item 1 Category 0	0	0	0	0	0	0	0	0	0	0
Item 1 Category 1	1	0	0	0	0	0	0	0	0	0
Item 2 Category 0	0	0	0	0	0	0	0	0	0	0
Item 2 Category 1	0	1	0	0	0	0	0	0	0	0
Item 3 Category 0	0	0	0	0	0	0	0	0	0	0
Item 3 Category 1	0	0	1	0	0	0	0	0	0	0
Item 4 Category 0	0	0	0	0	0	0	0	0	0	0
Item 4 Category 1	0	0	0	1	0	0	0	0	0	0
Item 5 Category 0	0	0	0	0	0	0	0	0	0	0
Item 5 Category 1	0	0	0	0	1	0	0	0	0	0
Item 6 Category 0	0	0	0	0	0	0	0	0	0	0
Item 6 Category 1	0	0	0	0	0	1	0	0	0	0
Item 7 Category 0	0	0	0	0	0	0	0	0	0	0
Item 7 Category 1	0	0	0	0	0	0	1	0	0	0
Item 8 Category 0	0	0	0	0	0	0	0	0	0	0
Item 8 Category 1	0	0	0	0	0	0	0	1	0	0
Item 9 Category 0	0	0	0	0	0	0	0	0	0	0
Item 9 Category 1	0	0	0	0	0	0	0	0	1	0
Item 10 Category 0	0	0	0	0	0	0	0	0	0	0
Item 10 Category 1	0	0	0	0	0	0	0	0	0	1

$\pm$ 2 logits and centred at 0. New case abilities and item difficulties were generated for each replication.

For each replication of data sets with associated case abilities and item difficulties, simulated data sets were produced with various induced values of ρ, namely for $\rho = 0.5,\ 1.0,\ 1.5,\ 2.0$ and 2.5 logits.

For each value of ρ, the infit (weighted mean square) and the User Defined infit statistics are presented in the following ways:

FIT A – For each individual item (in the normal manner) (Table 5.3)
FIT B – For each pair of dependent items (i.e. Item 1 & 2, Item 3 & 4, etc.) (Table 5.4)
FIT C – For two sets of 5 items, the first set contains the items which do not depend on another item (i.e. Item 1, 3, 5, 7 & 9) and the second set contains the items which do depend on another item (i.e. Item 2, 4, 6, 8, 10) (Table 5.5).

The results of the obtained infit values are presented in the figures below. The infit values shown were averaged over the 1,000 replications.

Figure 5.1 shows the mean infit values for each item (Fit A). It is clear that the individual item infit does not pick up any existence of the item dependence, as all the infit values are close to their expectation of 1. While there is a slight zigzag pattern, the values are too close to 1 to identify any misfit, even when the dependency level is highest ($\rho = 2.5$). The t-statistics for these fit values are all around 1 or lower and hence not statistically significant.

Table 5.4 Design matrix FIT B – Infit: pairs of dependent items

	Fit tests				
Items and Categories	F1	F2	F3	F4	F5
Item 1 Category 0	0	0	0	0	0
Item 1 Category 1	1	0	0	0	0
Item 2 Category 0	0	0	0	0	0
Item 2 Category 1	1	0	0	0	0
Item 3 Category 0	0	0	0	0	0
Item 3 Category 1	0	1	0	0	0
Item 4 Category 0	0	0	0	0	0
Item 4 Category 1	0	1	0	0	0
Item 5 Category 0	0	0	0	0	0
Item 5 Category 1	0	0	1	0	0
Item 6 Category 0	0	0	0	0	0
Item 6 Category 1	0	0	1	0	0
Item 7 Category 0	0	0	0	0	0
Item 7 Category 1	0	0	0	1	0
Item 8 Category 0	0	0	0	0	0
Item 8 Category 1	0	0	0	1	0
Item 9 Category 0	0	0	0	0	0
Item 9 Category 1	0	0	0	0	1
Item 10 Category 0	0	0	0	0	0
Item 10 Category 1	0	0	0	0	1

Table 5.5 Design matrix FIT C – Infit: set of dependent and independent items

	Fit tests	
Items and Categories	F1	F2
Item 1 Category 0	0	0
Item 1 Category 1	1	0
Item 2 Category 0	0	0
Item 2 Category 1	0	1
Item 3 Category 0	0	0
Item 3 Category 1	1	0
Item 4 Category 0	0	0
Item 4 Category 1	0	1
Item 5 Category 0	0	0
Item 5 Category 1	1	0
Item 6 Category 0	0	0
Item 6 Category 1	0	1
Item 7 Category 0	0	0
Item 7 Category 1	1	0
Item 8 Category 0	0	0
Item 8 Category 1	0	1
Item 9 Category 0	0	0
Item 9 Category 1	1	0
Item 10 Category 0	0	0
Item 10 Category 1	0	1

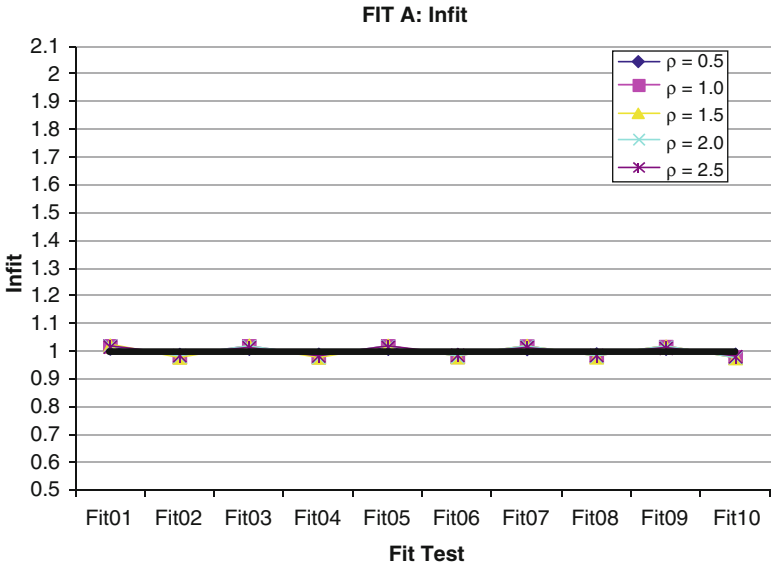


Fig. 5.1 FIT A – mean infit: individual items

Figure 5.2 shows the mean User Defined Fit statistics for the pairs of dependent items (Fit B). The figure shows that as the degree of dependence increases, the infit values increase. When $\rho = 0.5$ logits the infit values are around 1.15 with statistically significant t-statistics of around 3.7. When $\rho = 2.5$ logits the infit values are around 1.7 with t-statistics highly statistically significant at around 13.4.

Figure 5.3 shows the variances, over the 1,000 replications, for the User Defined Fit tests of the dependent pairs of items (Fit B). Although the variances are small, the figure shows that as the values of ρ increase, the variances of these User Defined Fit statistics increase.

Figure 5.4 shows the mean User Defined Fit Statistics for the five items with no induced dependence and the five items with induced dependence (Fit C). When the infit values are taken of the five items with no induced dependence and the five items with induced dependence, then as the degree of dependence increases, the infit values decrease. When there is extremely high dependence the scores of these two sets of five items compared to the total score discriminate a lot more than expected than predicted by the model.

5.3.3 Simulation 2: Testing Violations of Multi Dimensionality

So far the focus has been on violations of the model by inducing local dependence into the data. Andrich and Marais (2005) refer to this as 'response dependency'.

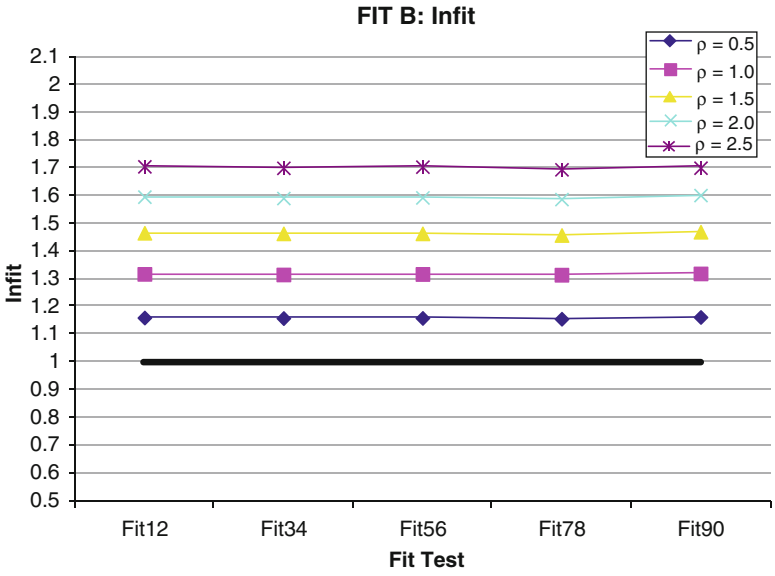


Fig. 5.2 FIT B – mean infit: pairs of dependent items

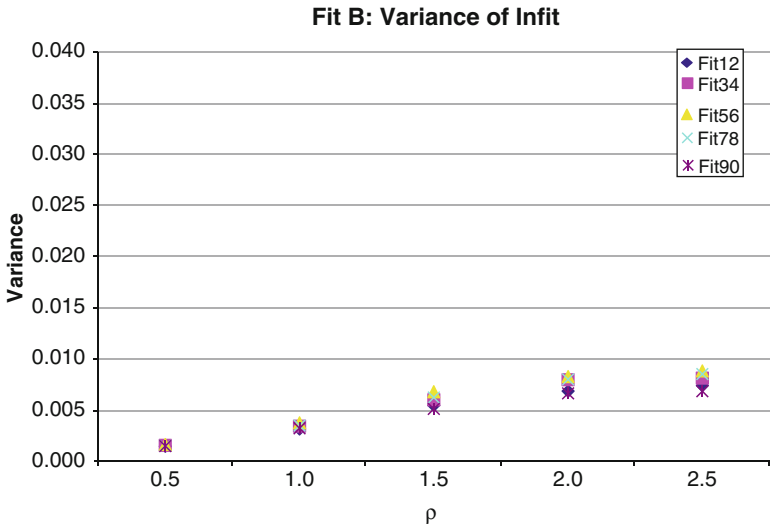


Fig. 5.3 FIT B – variance of infit: pairs of dependent items

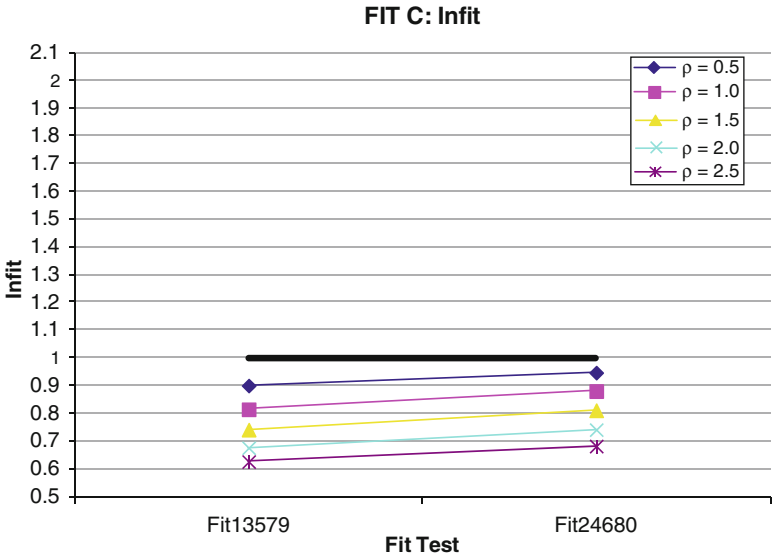


Fig. 5.4 FIT C – mean infit: set of dependent and independent items

Now the User Defined Fit Statistics are investigated when assumptions of uni-dimensionality are violated. This is referred to as 'trait dependency' by Andrich and Marais (2005).

Data sets were generated for 1,000 cases with dichotomous item 1–10 on dimension one and item 11–20 on dimension two. 1,000 replications were undertaken. Person abilities on both dimensions were generated from a uni variate normal distribution with mean 0, variance 1, and a correlation of 0.5 was set between the two dimensions. Item difficulties were generated from a uniform distribution with a range of $\pm$ 2 logits and centred at 0. For each replication, the generated data was modelled both as a one dimensional (1D) model and as a two dimensional (2D) model.

The infit (weighted mean square) and the User Defined infit statistics are presented in the following ways:

FIT A – For each individual item (in the normal manner)
FIT B – For each dimension (i.e. Items 1–10 & Items 11–20)

Figure 5.5 shows when the multi dimensionality is not accounted for (in a one dimensional model), the usual (averaged) infit values for each item do not pick this up. When this multi-dimensionality is accounted for (in a two dimensional model), the data is expected to fit the model and it looks like it does. In fact regardless whether the multi-dimensionality is or is not accounted for the individual infit values are close to their expectation of 1.

Figure 5.6 shows the (averaged) User Defined Fit Statistics for each dimension, again when the multi dimensionality is not accounted for (in a one dimensional

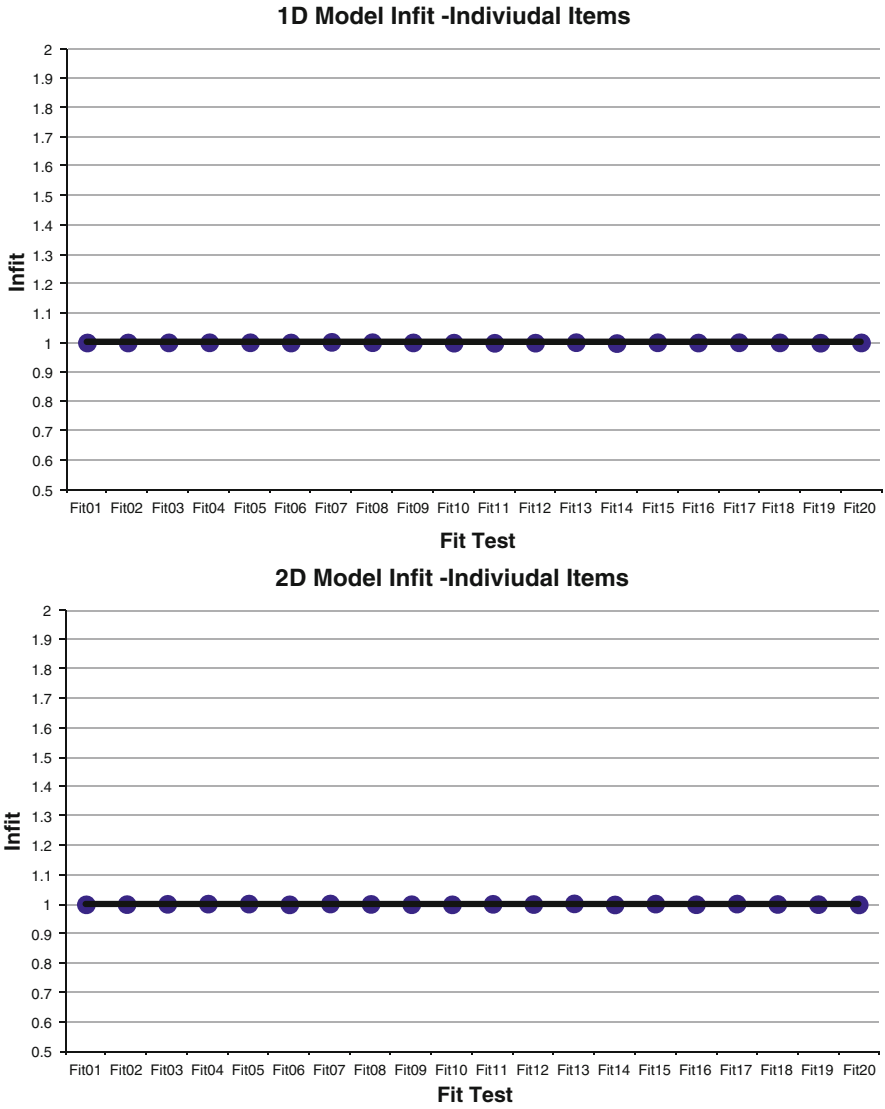


Fig. 5.5 FIT A – mean infit: individual items

model) and when it is accounted for (in a two dimensional model). Under the one dimensional model, the data is not expected to fit the model and clearly it does not. The infit values are around 1.4 with t-statistics of around 8. Under the two dimensional model, the data is expected to fit the model and clearly it does. The infit values are at their expectation 1 with t-statistics close to 0.

Simulation 2 shows the ability of the User Defined Fit Statistics to pick up violations of uni dimensionality in a uni dimensional model.

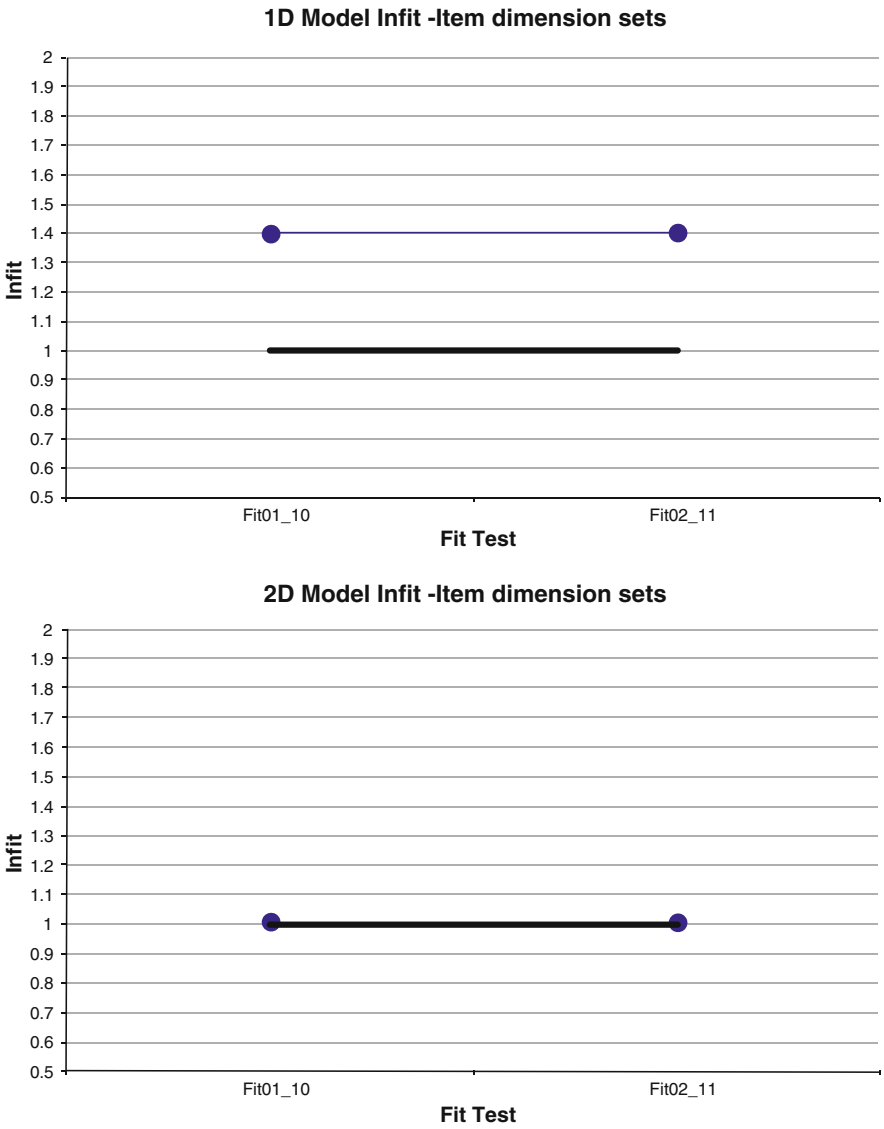


Fig. 5.6 FIT B – mean infit for each dimension

5.4 Analysis of Real Data

Two real data sets are also analysed, the first with an aim of identifying violations of local independence, the second with the aim of identifying violations of multi dimensionality.

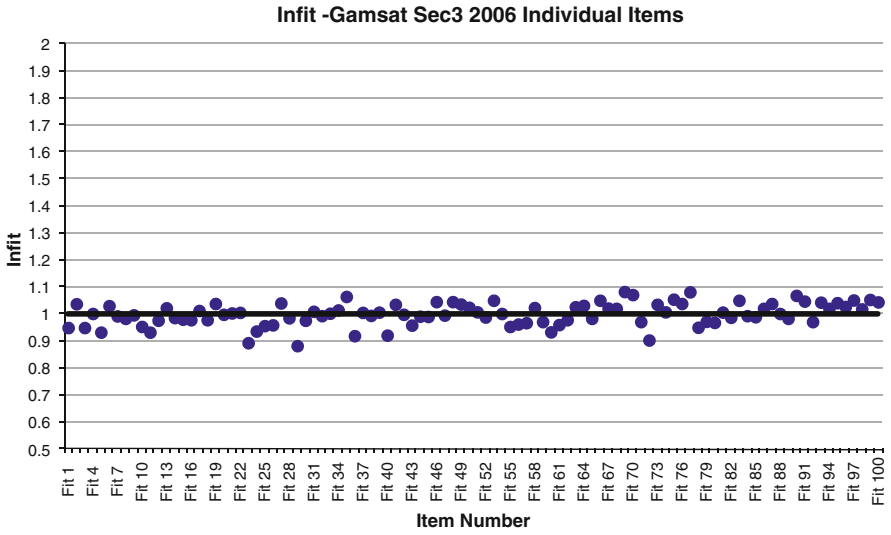


Fig. 5.7 FIT A – infit: individual items

5.4.1 Real Data Set 1: Section 3 of GAMSAT 2006

The first real data set investigated is section 3 of the Graduate Australian Medical Admissions Test (GAMSAT) for 2006. Section 3 is the Science Component of GAMSAT which included 4,232 cases and 100 items. These items are made up of 35 bundles of items. Each item bundle consists of a set of items relating to a common stimulus. The number of items in each item bundle ranged between 1 and 6 items.

The aim of this analysis is to investigate these item bundles to see if local dependence exists within them. Infit and User Defined infit statistics are presented for:

FIT A – For each individual item (in the normal manner)
FIT B – For each item bundle

Figure 5.7 shows the usual infit statistics are all close to their expectation of 1. These are the final infit values after the data was analysed and accounted for misfit and miss keys.

Figure 5.8 shows the User Defined Fit Statistics for each item bundle. This identifies some bundles of items which seem to display some evidence of dependence. The ten item bundles that show the most misfit are labelled in this figure.

Unfortunately, due to the secure nature of this test, the actual content of the items cannot be revealed to investigate the cause of item dependence within the item bundles further.

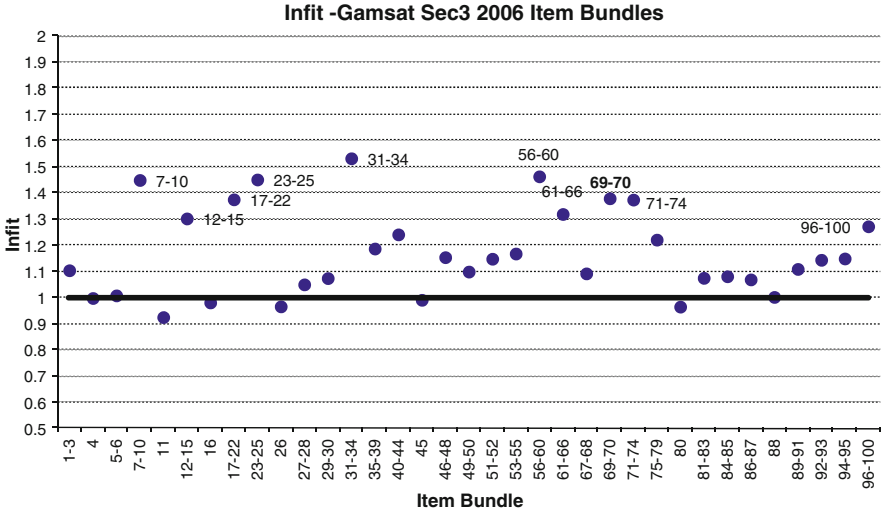


Fig. 5.8 FIT B – infit: pairs of dependent items

Table 5.6 Cross tabulation to investigate item dependency

q69 * q70 cross tabulation					
			q70		
			0	1	Total
q69	0	Count	1,631	553	2,184
		% within q69	74.7 %	25.3 %	100.0 %
		% within q70	63.1 %	33.6 %	51.6 %
	1	Count	953	1,095	2,048
		% within q69	46.5 %	53.5 %	100.0 %
		% within q70	36.9 %	66.4 %	48.4 %
Total		Count	2,584	1,648	4,232
		% within q69	61.1 %	38.9 %	100.0 %
		% within q70	100.0 %	100.0 %	100.0 %

What can be done however, is a cross tabulation of an item bundle to possible try to investigate the existence of item dependence further. Table 5.6 takes the 2 pair item bundle for item 69 and item 70 as an example.

The User Defined Fit value for this item bundle is 1.38. 48.4 % of candidates answered q69 correctly, while 38.9 % of candidates answered q70 correctly. Therefore the product of both item facilities is 18.8 %. However the actual proportion of candidates answering both items correctly is 25.9 %. The ratio of these percentages is 1.38. Under item independence, the ratio of the percentages would be expected to be close to 1. Although raw score percentages do not separate item difficulty from person ability, this method gives somewhat of a guide to the dependence between items.

5.4.2 *Real Data Set 2: Student Attitudes for the Victorian Learning Difficulties Project*

The second data set investigated is a questionnaire of Student Attitudes for the Victorian Learning Difficulties Project. The questionnaire consists of 12 rating scale items ranging from Strongly Agree, Agree, Disagree and Strongly Disagree. According to confirmatory factor analytic methods, this questionnaire contains 3 dimensions, with 4 items in each dimension. The data contained 2,279 responses.

The aim of this exercise is to confirm the existence of 3 dimensions using the User Defined Fit Statistics, by conducting an analysis of fit using a one dimensional (1D) and a three dimensional (3D) model and providing infit values for:

FIT A – For each individual item (in the normal manner)
FIT B – For each dimension

Figure 5.9 shows the individual item infit statistics. For the uni dimensional model, all are reasonably close to their expectation of 1 except for item 9. A similar result holds for the individual infit statistics when modelled with a three dimensional model.

The User Defined Fit Statistics for each dimension, when the multi dimensionality is not accounted for (in a one dimensional model) and when the multi dimensionality is accounted for (in a three dimensional model), are shown Fig. 5.10. Under the one dimensional model, the data is not expected to fit the model and clearly it does not, as the infit values are 1.8, 1.7 and 1.6 for the items in dimension one, two, and three respectively (with t-statistics of 21, 18 and 14 respectively). The User Defined Fit Statistics pick up the dimensionality violation. Under the three dimensional model, the data is expected to fit the model and clearly it does, as the infit values are at their expectation of 1 with t-statistics close to 0.

5.5 Concluding Observations

The User Defined Fit Statistics allow for the fit of linear combinations of items to be tested, using linear combinations of individual contributions to the sufficient statistics and comparing them to their expectations. That is, the fit of a set of items (as opposed to just individual items) can be tested, to see if such a set functions differently than predicted by the model.

After inducing violations of local item independence in simulated data, it was found that the User Defined Fit Statistics identify such violations of the model. After producing a two dimensional simulated data set, again, the User Defined Fit Statistics were able to identify such violations of the model as well. Two real data sets were also analysed. The presence of some item dependence among some item bundles in the GAMSAT 2006 Section 3 data was identified which requires some further investigation. As well as this, the existence of three dimensions in the Victorian LD Student Attitudes was confirmed using the User Defined Fit Statistics.

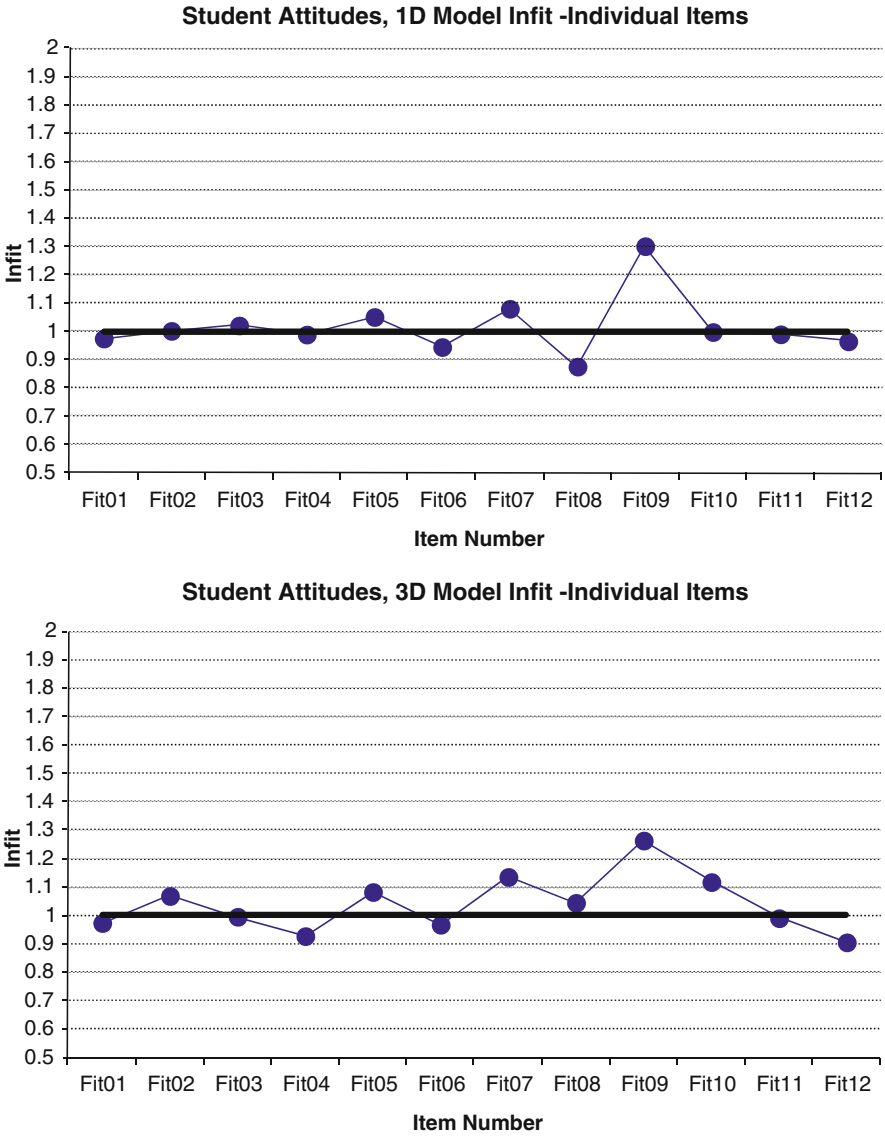


Fig. 5.9 FIT A – infit for individual items

One restriction of utilising the User Defined Fit Statics efficiently is the requirement of a priori knowledge of possible item combinations that may violate the model's assumptions. However, it is often the case in assessment scenarios that items are grouped in either bundles, sub scales or sub strands. This enables a priori assumptions about such item combinations to be made. Hence these User Defined Fit Statistics may be extremely useful in identifying violations of local independence and uni dimensionality in practice.

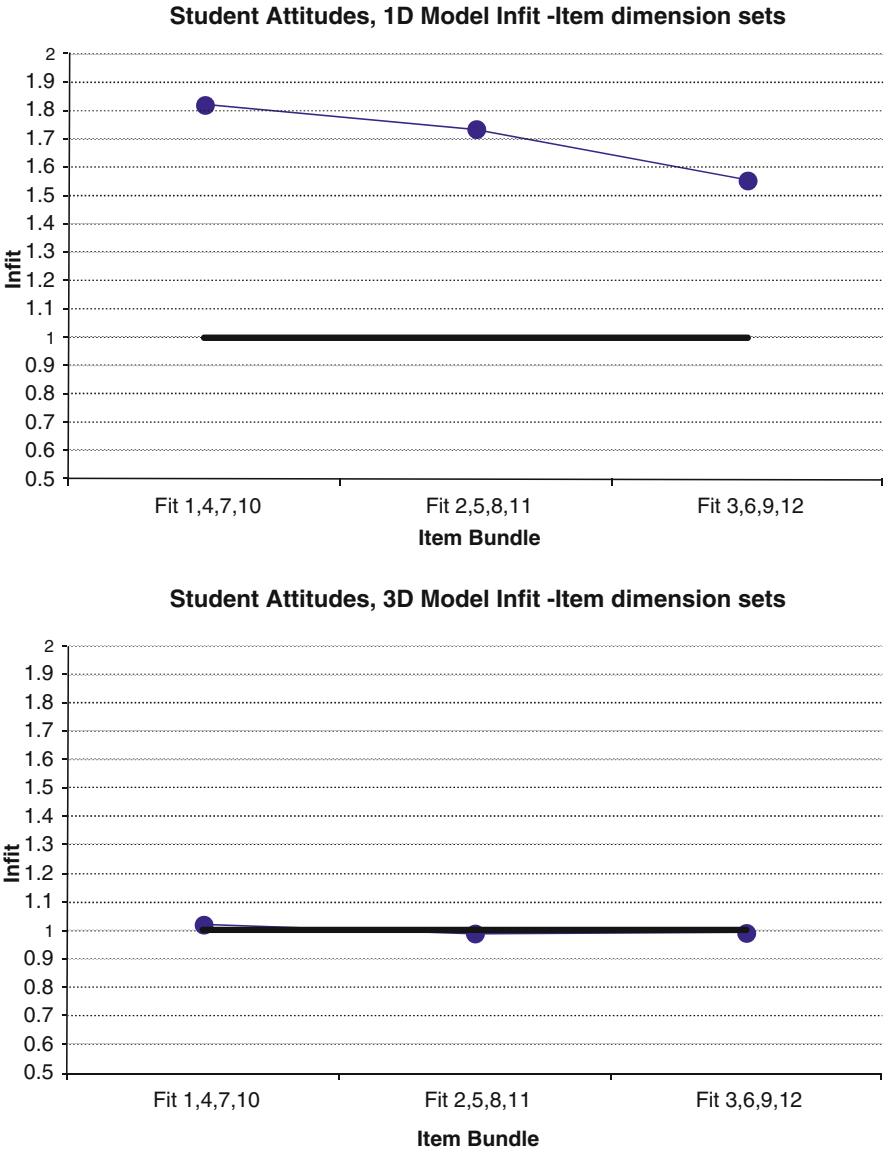


Fig. 5.10 FIT B – infit for each dimension

References

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficient multinomial logit. In G. Engelhard Jr. & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 143–166). Norwood: Ablex.

Adams, R. J., & Wu, M. L. (2009). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalised item response models. *Journal of Applied Measurement, 10*(4), 355–370.

Adams, R. J., Wilson, M. R., & Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioural Statistics, 22*(1), 47–76.

Andrich, D. A., & Marais, I. (2005). *Studies on the effect of violations of local independence on scale in Rasch models*. Funded by an ARC grant with industry partners MCEETYA, IIEP, and ACER.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *The British Journal of Mathematical and Statistical Psychology, 33*, 234–246.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESE Press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models*. Unpublished Master of Education Thesis, Melbourne University.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest version 2.0 [computer program]*. Camberwell: ACER Press, Australian Council for Educational Research.

Chapter 6
Investigating the Consequences of the Application of Formative Evaluation to Reading-Writing Model

Hong Yang, Hong Zhou, and Yan Zhao

Abstract Nowadays, we can see reading accounts for a large proportion in some large-scale English tests in China and abroad, such as CET-4, CET-6, TEM-4, GRE, IELTS, and TOEFL. Many studies show that a considerable number of Chinese students can get high marks but they have never developed true reading skills and one of the major reasons is that they lack adequate effective reading. How to combine reading and writing effectively in teaching has become an important research field in EFL teaching in China. This study adopts a comparative way, in which two freshman English major classes with similar motivation, proficiency, sex distribution and taught by the same teacher will be engaged. Class A, where formative evaluation is practiced, is the experimental group and Class B with the conventional assessment is the control group. The data elicited from interviews, observation and students' portfolios are used for qualitative analysis and the data from questionnaire and reading tests are used for quantitative analysis by using the software SPSS 16.0. The results thus obtained show that the application of formative evaluation to reading-writing model is of great significance in anchoring learners' identity and improving students' self-management of learning in that it is the students who can tell the teacher what they have learned and what they are learning.

Keywords Formative evaluation • Reading-writing model • Consequences • Writing through reading • Reading through writing

---

This project is financially supported by the Humanities and SocialScience Department, Ministry of Education, P. R. China during 2009–2012 (project code: 09YJA740054) and by the Education Department of Zhejiang Province during 2009–2012 (project code: yb09064).

H. Yang (✉) • H. Zhou
English Department, Faculty of Foreign Studies, University of Jiaxing,
Jiaxing, Zhejiang Province, China
e-mail: yanghong_158@yahoo.com.cn

Y. Zhao
Department of Curriculum & Instruction, College of Education,
University of Alabama, Tuscaloosa, AL, USA

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_6,

6.1 Introduction

As is known to all that it is the essential goal in EFL classroom to develop students' intercultural competence, which includes knowledge, skills, abilities and faculty. Among the four skills, listening and reading belong to receptive skills, and speaking and writing belong to productive skills. Generally speaking, if learners get more receptive skills, their productive skills can be greatly improved. And Many studies have proved that more reading can improve writing in many ways (Krashen 1989; Janopoulous 1986; Johns 1993; MeMesser 1997; Harmer 2000; Esmaeili 2000). In recent years, some Chinese scholars have tried combining reading and writing in their writing or reading classroom (Yang and Dong 2010; Xu and Gao 2007; Li and Wang 2011), but their focus is on the impacts of reading on writing. Especially Xu & Gao found that there was no significant impact of writing on reading. And the truth is that they tried to combine the writing syllabus with reading one, and the samples used in writing class were the materials used in reading class and the topics for writing assignments were chosen from reading textbooks, "ignoring the application and transferring of the knowledge and abilities of writing to reading in turn". Up to now, there is still a serious need for further empirical studies in this area in China.

In the reading-writing teaching model, how to keep "reading through writing" and "writing through reading" in balance to improve the students' integrated use of reading and writing abilities? What effective mechanism can be established to better supervise and monitor students' reading learning? Based on Bloom's taxonomy of educational objectives (revised) and formative evaluation theories, the present study provided a tentative reading-writing model with formative evaluation taken into consideration, and focused on whether and how the model promoted EFL reading development of Chinese college English majors. The consequences of the model were investigated in terms of learners' perception of formative evaluation, their reading process (strategy use) and reading comprehension. The present study would be of great significance in that it provides some implications for EFL teaching and learning.

The rest of this chapter is organized as follows: in Sect. 6.2, the rationale of the present study is briefly introduced; then the methodology is presented in Sect. 6.3; and the results and discussion are presented in Sect. 6.4; and the conclusions are summarized in the last section.

6.2 The Rationale of the Present Study

6.2.1 The Revised Bloom's Taxonomy of Educational Objectives

Bloom's Taxonomy is a classification of learning objectives within education proposed in 1956 by a committee of educators chaired by Benjamin Bloom who also edited the first volume of the standard text, *Taxonomy of educational*

objectives: the classification of educational goals (Bloom et al. 1956). And to keep the importance of Bloom's work relative to today's theories, Lorin W. Anderson, David R. Krathwohl, et al. (2001) revised Bloom's original taxonomy in *A Taxonomy for Learning, Teaching and Assessing – A Revision of Bloom's Taxonomy of Educational Objectives* (hereafter referred to the revised edition). Although Bloom's taxonomy of educational objectives has been widely accepted by the people and produced far-reaching effect on the United States education of last century, the revised edition has its obvious advantages by combining both the cognitive process and knowledge dimensions, paying more attention to the constructivism learning (i.e. meaningful learning). The revised Bloom's taxonomy takes the form of a two-dimensional table, called "Taxonomy Table". The Knowledge Dimension, on the left side, is composed of four levels that are defined as Factual, Conceptual, Procedural, and Meta-Cognitive. The Cognitive Process Dimension, across the top of the grid, consists of six levels that are defined as Remember, Understand, Apply, Analyze, Evaluate, and Create. Therefore, the revised taxonomy provides a very powerful tool to the current study, helping the researchers identify and determine reasonable educational goals, conduct proper teaching and design effective assessment tasks.

According to *The Syllabus for Undergraduate English Majors in Higher Educational Institutions* (2000), there are four extensive reading programs. The instructional design of reading-writing model was based on Program 2 in this present study, with all educational goals, activities and assessment tasks put in Table 6.7 (see Appendix A). The only distinguishing factor – formative evaluation was taken into consideration for the experimental group in the whole process of teaching while a posttest was conducted for both the two groups for a comparison at the end of the semester.

6.2.2 Formative Evaluation Theory

The concept, formative assessment (hereafter using formative evaluation), was first proposed by M. Scriven in 1967. The American educationist Bloom applied it to practice in educational assessment. Afterwards, a great deal of researches have been conducted and great achievements attained at home and abroad.

Although scholars use different definitions, their views have something in common: relative to summative evaluation, formative evaluation happens in teaching and learning process and the purpose of collecting feedback is to improve teaching and enhance learning (Byrant et al. 2002; Tian 2003). Biggs and Watkins (2001) think that formative evaluation can help students find out their problems as well as the solutions to the problems in learning through the interaction between teachers and students. And such tools as observation, interview, questionnaire, test and portfolio are often used for formative evaluation (Wang 2007). However, it should be pointed out that the different tools have different functions in some specific situations. In this present study, the data elicited from questionnaire, classroom observation and students' portfolios are used for investigating students'

learning strategies and learning experience; the data from interview for investigating students' perception and attitudes towards formative evaluation; and the data from reading tests for investigating students' reading proficiency (a construct).

6.2.3 Previous Studies

A case study by Chen Kai (2002) shows that writing can play a guiding role for extracurricular reading. He believes that the reading tasks set by the teachers can activate the students' reading motivation in that an urgent need for seeking answers enable them to have a strong desire for knowledge. Generally speaking, there are two ways for students to complete their reading tasks: oral and written forms. The inside and outside activities concerning oral form may include discussions, debates, lectures, talking about reading experiences and retelling while those concerning written form include book report, listing some details, abridging, summarizing and filling the blanks to complete a passage. Compared with oral form, the advantages of written one lies in two aspects. On the one hand, it's not easy for students to avoid finishing their writing tasks which need the participation of all the students. On the other hand, writing tasks can enable students to have plenty of time for carrying out a wide range of reading and seeking the answers by analyzing, synthesizing and generalizing all the information they have got. The statistical results of this study show that the ability of reading humanity and science articles of the students from Class A (experimental group) were improved significantly.

And another case by Li Meiying and Lv Qin (2002) also show that the task of writing book report can be an effective means for supervising and monitoring the students' extracurricular reading in reading-writing model, with the book report taken into peacetime achievements of 10 % as a guarantee. Up to now, there are still very few overall empirical research on the effects of writing on reading.

6.3 Research Methodology

6.3.1 Research Questions

The current study is to explore the effectiveness of the application of formative evaluation to the reading-writing model in educational system of mainland China. The key research question is: How does the new model influence Chinese undergraduates' English reading development? And the following three sub-questions will be focused on so as to answer the key question:

1. Are there significant difference between the experimental class (Class A) and the non-experimental class (Class B) in terms of their EFL reading strategies at the end of the experimental semester?

2. Are there significant differences between the experimental class (Class A) and the non- experimental class (Class B) in terms of their EFL reading proficiency at the end of the experimental semester?
3. How do the students in the experimental perceive formative evaluation applied to the reading-writing model in the course of learning?

6.3.2 Participants

This study adopted a comparative way, in which two freshman English major classes in Jiaxing University with similar motivation, proficiency, sex distribution and taught by the same teacher were engaged. Class A, where the new model was practiced, is the experimental group and Class B with the conventional assessment is the control group. Each class consisted of 33 students aging from 18 to 21. Most of them had been learning English for more than 8 years. Students in Class A were divided into three groups (high-, middle-, and low- proficiency) on the basis of their reading proficiency and two students with similar proficiency were selected from each group for the interviews.

6.3.3 Instruments

Creswell (1994) states that combination of both quantitative and qualitative methods is helpful to reveal the holistic, naturalistic and inductive aspects of the phenomena under investigation. The present study adopts a multi-method design, with data sources included from questionnaire, classroom observation, interviews with the selected students, tests, and students' essays, book reports, journals and reflective essays in their portfolios.

6.3.4 Procedures

In order to answer the research questions, the data was collected in three phases: (1) Before the reading-writing model project: both classes took a reading test (pretest) as a diagnostic examination before the first lecture, with the students' reading skills focused on, such as identifying the main idea and other salient features in a text, generalizing and drawing conclusions, understanding some implied meanings, and so on. (2) During the project: In the second semester of 2009–2010 academic year, the researchers read the journals, essays and reflective essays in the six students' portfolios in the middle of the semester and at the end of the semester. (3) After the project: the two classes took another test (posttest). The data from interviews were transcribed and put into computer. All the students in the two groups were required to respond to the strategy questionnaire. The data were collected and stored in Excel.

6.3.5 *Data Analysis*

As mentioned above, the present study is a multi-method design, in which quantitative and qualitative data were collected to be used as evidence to answer the research questions. All the quantitative data were analyzed by using the Statistical Package for Social Science (SPSS), version 16.0 for windows. The strategy questionnaire issued to the two classes were carried out and the students' responses to the items were collected and reorganized. All the test items in the questionnaire were presented using a five-point Likert scale. The scores from the statements were written in the following formulas: 1 = 5; 2 = 4; 3 = 3; 4 = 2; 5 = 1. Comparison of means was conducted to see if there were significance differences between the two classes at the end of experiment. Comparisons were analyzed with two independent samples *t* test with α set at .05. Moreover, the paired *t* test was applied to see any change of reading comprehension. The collected data from teachers' observation, interviews, journals and reflective essays in the six students' portfolios were organized and categorized, and then sorted for themes and patterns.

6.4 The Results and Discussion

6.4.1 *Questionnaire*

All participants were required to respond to the questionnaire on reading strategy. Data from this instrument were stored in Excel and SPSS was used to analyze the differences between the two groups.

Questionnaire (See Appendix B) is composed of 15 items concerning students' reading strategies, which were adopted from Zhang (2005), with some items modified according to the educational objectives for the participants. Table 6.1 is the general description of the statistical results of the strategy questionnaire. All the students (N = 66) completed the survey. Significant difference was found in reading strategy between Class A and Class B (t = 7.375, p = .000).

Comparison of equality of means was conducted item by item. And the detailed results of independent samples *t* tests of reading strategy for each item in the questionnaire were summarized and presented in Table 6.2 as follows:

Strategy item 1 is concerned with Objective 2 (See Table 6.7, Appendix A) – vocabulary enlarging and remembering. Students in Class A used this strategy more frequently than those in Class B, which could be inferred from statistical results in Table 6.2: t = 2.766, p = .008. Students should grasp 4,000–4,500 words by the end of the second semester and word spelling is very important for their writing. When the students wrote book reports or reviews, they returned to the reading materials for getting key words or phrases, thus they went over and remembered the words consciously and unconsciously.

Table 6.1 General description of independent samples *t* test of reading strategy

Class	N	Mean	Std. deviation	t	df	Sig. (2-tailed)
A	33	63.33	5.152	7.375	64	.000
B	33	49.67	9.316	7.375	49.89	.000

Table 6.2 A summary of independent samples *t* tests of reading strategy

Item	*t*	*p*
Strategy 1	2.766	.008
Strategy 2	4.828	.000
Strategy 3	3.065	.003
Strategy 4	4.735	.000
Strategy 5	2.862	.006
Strategy 6	2.766	.008
Strategy 7	4.828	.000
Strategy 8	3.861	.000
Strategy 9	3.766	.000
Strategy 10	4.189	.000
Strategy 11	3.465	.001
Strategy 12	3.762	.000
Strategy 13	4.079	.000
Strategy 14	5.349	.000
Strategy 15	1.731	.088

Strategy items 2 and 3 were related to the use of grammatical and translation knowledge. It seems that students are not encouraged to stop often to analyze the sentence structures or translate English sentences into Chinese, but analyzing long and difficult sentences can help students comprehend the sentences and use the sentence patterns in their writing. And translating, in fact the use of native language, can effect positive transfer for learners, especially for freshmen. In the present study, students in Class A tended to use Strategy items 2 and 3 more frequently (Strategy 2: $t = 4.828$, $p = .000$; Strategy 3: $t = 3.065$, $p = .003$). There were significant differences between Class A and Class B.

Strategy items 4 and 5 focused on students' treatment of new words. According to Objective 3, students are able to take use of some skills and strategies to improve their reading speed (70–120 words/min). In order to reach this goal, the participants were required to skim and scan the whole reading material for getting the general idea, not looking up the new words in a dictionary but guessing instead when they were reading. Students in Class A tended to use Strategy items 4 and 5 more frequently (Strategy 4: Mean of Class A = 4.52, Mean of Class B = 3.24; Strategy 5: Mean of Class A = 4.42, Mean of Class B = 3.73).

Strategy items 6 and 7 were related to the strategies of reading in depth. Students were instructed to distinguish key parts or targeted parts from unimportant ones and asked to read slowly word by word and repeatedly, which could help them improve their comprehending of the main idea. There were significant differences

between ClassA and Class B (Strategy 6: t = 2.766, p = .008; Strategy 7: t = 4.828, p = .000).

Strategy items 8 and 9 were involved in Objective 4 (Analyzing). According to the revised Bloom's taxonomy, "analyzing" means learners can distinguish between the different parts, recognize the relationship between parts and the relationship between parts and the whole textual structure or the overall purpose of materials, including differentiating, organizing and attributing. In the present study, students were asked to take use of conceptual knowledge (theme and genres) to analyze the textual structure and arrive at a conclusion by using logic and reasoning. And they could benefit a lot from analyzing when writing a passage in the form of some genre. Students in Class A tended to use Strategy items 8 and 9 more frequently (Strategy 8: Mean of Class A = 4.24, Mean of Class B = 3.27; Strategy 9: Mean of Class A = 4.36, Mean of Class B = 3.42).

Strategy items 10, 11, 12, 13 and 14 were concerned with Objective 3 (applying of meta-cognitive knowledge) and Objective 6 (Creating). Students in both Class A and Class B were instructed with meta-cognitive knowledge, such as strategic knowledge, knowledge about cognitive tasks and self-knowledge, however, students in Class A seemed to adopt all these strategies more frequently in that such tools as self-assessment and peer-assessment were introduced into their learning to keep reading(input) and writing(output) in balance. *t* test showed significant differences between the two groups for each item (Strategy 10: t = 4.189, p = .000; Strategy 11: t = 3.465, p = .001; Strategy 12: t = 3.762, p = .000; Strategy13: t = 4.079, p = .000; Strategy14: t = 5.349, p = .000).

As for Strategy item 15, "I turn to my teachers or classmates for help when I have some difficulty in comprehending", *t* test showed no significant difference between the two groups.

To sum up, as is shown in Table 6.2, there were significant differences between the two groups in all the items except Item 15 (t = 1.731, p = .088 > .05). It indicated that Class A could take use of strategies in item 1–14 more frequently to improve their reading and writing than Class B.

6.4.2 Pre-test and Post-test

A pre-test was conducted to verify if there was any significant difference in English comprehension between Class A and Class B. Then, at the end of the semester, a post-test was also conducted to see if students in Class A improved their reading ability through writing more effectively. Both the pre-test paper and post-test paper are the reading parts of CET4 (fast reading and reading in depth) in that they have high reliability and validity. All the participants were required to complete fast reading within 15 min and reading in-depth within 35 min. Data collected were analyzed by SPSS software. The results of independent samples *t* test were presented as follows.

Table 6.3 Descriptive statistics of pretest on Class A and Class B

Class	N	Mean	Std. deviation	Std. error mean
A	33	10.58	2.926	.509
B	33	11.31	3.306	.584

From Table 6.3, we can see that the average scores of Class A and Class B were quite similar, 10.58 and 11.31 respectively. Table 6.4 shows that there was no significant difference between Class A and Class B in reading comprehension at the very beginning of the experiment ($t = -.952$, $p = .345 > .05$).

By comparing the average scores of Tables 6.5 and 6.6, students of two groups both made progress in reading comprehension after a semester's learning. However, it seemed that students in Class A made greater progress than those in Class B, with their average scores being 15.33 and 12.52 respectively. Table 6.6 indicates that there were significant differences between Class A and Class B in reading comprehension ($t = 3.697$, $p = .000 < .05$).

6.4.3 Portfolios

Portfolios in this present experiment were used for the purpose of formative evaluation, with such records of students' learning included as journals, reflective essays, book reports, the original drafts, revised and final ones. Students were required to finish some task-based writing assignments after getting further and detailed reading of the materials in their textbooks, and to write book reports and reflective essays, and keep journals after they chose to read extracurricular materials.

At the middle of the semester, the researchers learned from the reflective essays that the six subjects were skeptical at the very beginning and they hoped to get more feedback and help from their teacher in adjusting their learning plan and choosing reading materials. However, at the end of the semester, they learned how to use some reading skills to get the information they needed for their writing effectively and to manage their own learning. The following is a part of one reflective essay:

> To tell the truth, I didn't expect to benefit much from this new model at the very beginning. However, I made a plan for the whole semester with the help of my teacher. For Unit 1, the teacher didn't ask us to finish the multiple-choice exercises as usual after reading but introduced the background and gave us instructions on how to read such articles. Then, we were assigned a writing task with detailed questions as guiding. I found that those questions were closely related to comprehending the main idea. I tried to take use of such reading skills as skimming and scanning, guessing the meanings of words or sentences, analyzing long and difficult sentences or reading the important part repeatedly. Through text-revision, I could grasp the major points of the text and use the words and phrases I had just learned to express myself in my writing! (reflective essay/03/12/2010)

There were significant differences in all the six subjects' final writing products compared with their revised drafts, in which comments, suggestions and opinions from their teacher and classmates as feedback were always found. Five of six

Table 6.4 Independent samples test of pre-test on Class A and Class B

		Levene's test for equality of variances		t-test for equality of means						
									95 % confidence interval of the difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std. error difference	Lower	Upper
Pretest score	Equal variances assumed	1.363	.247	−.952	63	.345	−.737	.774	−2.283	.810
	Equal variances not assumed			−.950	61.569	.346	−.737	.775	−2.287	.813

Table 6.5 Descriptive statistics of post-test on Class A and Class B

Class	N	Mean	Std. deviation	Std. error mean
A	33	15.33	3.017	.525
B	33	12.52	3.173	.552

subjects finished their book reports or reviews as required at the end of the semester. The dates in which they read the books and the names of the books were kept in their journals.

6.4.4 Interviews

Three semi-constructed interviews were conducted to investigate students' perception of formative evaluation applied to reading-writing learning at the beginning, middle and the end of the project respectively. The detailed questions to the six students are presented in Appendix C. The data from interviews were transcribed and stored in Software word file.

At the beginning of the project: The first three questions concerning the students' EFL reading and writing learning in their first semester. According to the six subjects, their teacher also adopted reading-writing model teaching, i.e. students were given the assignments of extracurricular reading and writing. However, there were not much additional instructions and feedback provided for their learning, they did not know what to read and how to read, as a result, they gradually lost their interest in reading and writing. And they were not satisfied with their learning except one from the high-proficiency group. As for Q4 and Q5, all the six subjects knew nothing about formative evaluation and would be willing to take part in the new reading-writing model activities.

At the middle of the project: It was expected that students in both Class A and Class B took part in the similar instructional activities except formative evaluation. The first two questions were concerning the teacher's constant instructions on the new reading-writing model. The six subjects talked to the teacher after they had carefully read the guidelines the teacher gave them. Even though they knew the process of the new model well, only four of them (two from high-proficiency group and two from middle-proficiency group) followed each step of the whole process of task-based writing: reading as required → writing → assessment (self-, peer-, or teacher assessment) → reading more → revising frequently → final draft. They thought that their reading and writing were greatly improved in that they were always more concerned about how to revise their essays according to others' opinions. But the two subjects from low-proficiency group complained that the process of assessment was too complicated and they were afraid of being laughed at by the other classmates because of their poor writing, instead, they always checked their essays based on criteria. They would rather read the materials repeatedly and were more familiar with the words and expressions. Accordingly, they felt their reading was improved more greatly. By far, to the six subjects, the new model activities appeared to be effective in facilitating the internalization, storage, or use

Table 6.6 Independent samples test of post-test on Class A and Class B

		Levene's test for equality of variances		t-test for equality of means						
									95 % confidence interval of the difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std. error difference	Lower	Upper
Posttest score	Equal variances assumed	1.129	.292	3.697	64	.000	2.818	.762	1.295	4.341
	Equal variances not assumed			3.697	63.838	.000	2.818	.762	1.295	4.341

of the new language. However, they needed more training and more chances to express themselves in class and after class.

At the end of the project: All the six subjects thought that the new model could improve their reading and writing to a greater extent than before, being effective in facilitating the internalization, storage, or use of the new language. One subject from the high-proficiency group said, "Formative evaluation is not a new term for me now. I know it can help me in many ways. One of the most important things I have realized is that I have learned how to use the different kinds of feedbacks from my teacher, my classmates and myself. When my shortcomings and mistakes were identified, I made great efforts to correct them. And my teacher helped me in analyzing my learning needs and giving suggestions on my study plan for next step learning several times during the semester. I read more original books in that I took use of some reading strategies to increase my reading speed considerably. I benefited a lot."

For the two subjects from low-proficiency group, the new model means more. They felt their motivation and confidence were greatly enhanced because final exam was not the only means to be used for their learning evaluation. One of them said, "I was given the right to choose the pieces of writing which I thought the most satisfying and to put them into my portfolio. My teacher noticed my progress in learning, which was a great encouragement for me. When I wrote my reflective essays, I always asked myself such questions: 'Did I use the right reading strategies and skills when reading? Did I use the right style when writing? Why (not) did I revise my essays according to the feedbacks from the teacher and my classmates? . . .. . .'. My portfolio is not just a collection of my essays, journals, book reports, and reflective essays, it is the record of my language development. Therefore, I think it is very useful."

As for Q6, all the six subjects gave comments and opinions from their own points of view. On the whole, they thought formative evaluation could promote their learning. On the one hand, formative evaluation functioned as a monitoring mechanism of their extracurricular reading and writing, with the individualization, independent learning strengthened. On the other hand, different feedbacks from the teacher and other classmates made them realize the significance of interactivity in language learning. They hoped that they would benefit more from this model in the future.

6.5 Conclusions

Above all, this present study adopted a multi-method design, with quantitative and qualitative data collected and analyzed. The results from questionnaire, post-test, portfolios and interviews showed that the new reading-writing model made a great impact on students' English reading development. It mainly behaves as follows:

Reading strategies & independent learning: The questionnaire was used to answer the first sub-question and the results showed that students in experimental

class (Class A) paid more attention to reading strategies when reading than those in the non-experimental class (Class B) at the end of the experimental semester. From Table 6.7 (see Appendix A), we can see two of the objectives (Objective 4 and 7) is concerning the "Applying" and "Evaluating" of "Meta-cognitive Knowledge". According to Lorin (2009), meta-cognitive knowledge includes strategic knowledge, knowledge about cognitive tasks and self-knowledge. In the process of teaching, all the knowledge was instructed by the teacher and the applying of the knowledge was assessed.

Strategic knowledge includes the knowledge about learning, thinking and solving problems as well as the knowledge for planning, monitoring and adjusting of cognition. In this present study, open and challenging tasks were provided for students to experience learning, think hard and solve the problems. Students were encouraged to produce learning outputs at different levels and write about how tasks had been done in their reflective essays. The knowledge about cognitive tasks required students should know when and how to use the strategic knowledge, i.e. using different "tools" to solve different problems. And self-knowledge appears particularly important because it is closely related to one's self-consciousness of weaknesses and strengths in cognition and learning, the range and depth of basic knowledge, and motivation and beliefs. The teacher, in this case, did not arrange everything but just provided students with help for diagnosing, analyzing their learning needs and "negotiating" with them about their learning goals so as to improve their learning. Therefore, students were led to independent learning, with self-identification, self-selection, self-assessment and self-adjustment concerned. This is also what modern education needs: the ultimate goal of teaching is to make students learn how to learn.

Learning goals: The overall goal of extensive reading is to improve students' reading comprehension. In this present study, the goals included getting the main idea, reading speed, vocabulary, analyzing and inducing abilities, reading strategies and writing ability (see Table 6.7, Appendix A). However, because there are many individual differences (age, aptitude, motivation, attitude, personality, cognitive style, learning strategies, and other factors) between foreign language learners (Freeman and Long 2000), students were allowed to have their own learning goals, selecting their own reading materials after class and making decisions for further study by "negotiating" with the teacher. Eventually, they could achieve the instructional goals.

Reading through writing: Even though writing belongs to productive skill, in Table 6.7, Objective 8 is "Creating" of "Factual Knowledge" and Objective 9 is "Creating" of "Conceptual Knowledge", that is to say, all the writing activities are related to reading tasks, As Wang (2007) and Chen (2002) state, the ultimate goal of reading is not just to recognize information in materials, but to express oneself by spoken or written language, and writing has a lot superiority. The results elicited from students' reflective essays and interviews showed that task-based writing helped students to decide what to read and how to read. When the students got different

feedbacks from the teacher and other classmates, they would constantly read the materials and revise their writing, with such knowledge as grammar, vocabulary, sentence patterns or style of writing reviewed, practiced and then consolidated, which, in turn, helped students to increase their reading speed and promote their reading comprehension. Formative evaluation is an effective mechanism of monitoring, diagnosing and guiding for teaching and learning (Yang and Zhou 2013).

As formative evaluation was used in an appropriate way, a win-win objective was achieved in this study. No doubt, formative evaluation also exerted positive effects on teaching. The teacher took use of the feedback to improve her teaching methods and achieve the final goals by repeating "evaluation—goal setting—teaching".

To summarize, there are some good implications for EFL teaching and learning, but there are also some limitations. One of them is that experiential learning touches upon affective dimension, which is an indispensable element, but not covered in this study, waiting for future research.

Appendix A

Table 6.7 The classification of objectives, activities and assessment tasks in the Taxonomy Table

	Cognitive process dimension					
Knowledge dimension	1. Remembering	2. Understanding	3. Applying	4. Analyzing	5. Evaluating	6. Creating
A. Factual knowledge	OB 1 Activities FA					OB 8 Activities FA
B. Conceptual knowledge		OB 2 Activities FA		OB 5 Activities	OB 6 Activities FA	OB 9 Activities FA
C. Procedural knowledge			OB 3 Activities FA			·
D. Meta-cognitive knowledge			OB 4 Activities FA		OB 7 FA	

Notes: *FA* formative evaluation; *OB* objective

Appendix B. Questionnaire on Reading Strategy

Part A Background Information

Name: _____ Gender: _____ Age: _____ Score of English in the Entrance Examination: _____

The purpose of Learning English: _____

Part B Below are 15 statements about people usually do in EFL reading and writing learning. There are no right or wrong answers to these statements. Please show the degree to which each statement applies to your actual reading by circling the number.

1= always 2= often 3=sometimes 4=seldom 5=never

Reading strategies	Descriptions	scale
1. Input of vocabulary	I spend longer time to grasp English word spelling and word formations as many as possible to clear the way in reading and writing.	1------2-----3------4-----5
2. Grammar	I analyze the difficult or long sentences by using some grammatical knowledge when reading.	1------2-----3------4-----5
3. Translation	I comprehend indirectly the reading materials by translating them into Chinese when directly processing the target language is not available.	1------2-----3------4-----5
4. The use of dictionary	I look up the key words if necessary after reading the whole article.	1------2-----3------4-----5
5. Guessing	I always take good use of discourse environment to guess the meanings of new words , sentences or new information	1------2-----3------4-----5
6. Reading word by word	I always give a study reading to the targeted part word by word.	1------2-----3------4-----5
7. Reading Repeatedly	I read repeatedly the key and difficult points selected after skimming.	1------2-----3------4-----5
8. Using logic	I tease apart the logical relations according to the contextual details.	1------2-----3------4-----5
9. Using reasoning	I come to a conclusion by judging, inferring and extending the implied meaning according to the information or clues provided.	1------2-----3------4-----5
10. Difficulty avoidance	I choose the reading materials which are moderately difficult for me. For example, I read the abridged versions of some original books.	1------2-----3------4-----5
11. Reading extensively	I take good use of any reading media to read extensively, such as original books, newspapers, magazines, electronic texts, etc..	1------2-----3------4-----5
12. self-management	I make the reading plans as required, with self-assessment and self-adjustment in the course of learning.	1------2-----3------4-----5
13. Social mediation	I participate in classroom activities and group activities after class.	1------2-----3------4-----5
14. Input and output of language	I write book reports, or book reviews and take notes in the course of or after reading.	1------2-----3------4-----5
15. External environment	I turn to my teachers or classmates for help when I get some difficult sentences or articles to comprehend.	1------2-----3------4-----5

Appendix C. Semi-Constructed Interview Questions

At the Beginning of the Project

Q1. How did your teacher deal with EFL reading in your first semester?
Q2. How did you deal with your EFL reading and writing in the first semester?
Q3. Were your reading and writing improved effectively in the first semester?
Q4. How much do you know about formative evaluation?
Q5. Are you willing to take part in the new reading-writing model activities?

At the Middle of the Project

Q1. Did your teacher always give you instructions to your reading and writing?
Q2. Are you clear about the process of the new reading-writing model? If yes, what are the steps?
Q3. Did your teacher often ask you to conduct self-assessment and peer-assessment?
Q4. What kinds of feedback could you get from teacher assessment, peer-assessment and self-assessment?
Q5. What do you think of the new reading-writing model activities so far?

At the End of the Project

Q1. Do you think the new model can help you improve your reading and writing?
Q2. In what ways do you think formative evaluation can help you improve your learning?
Q3. What kinds of feedback could you get from teacher assessment, peer-assessment and self-assessment?
Q4. How do you think of your portfolio? What are the advantages and disadvantages of portfolio?
Q5. Did you gain some confidence in your EFL learning?
Q6. Is there something else you would like to say about the new reading-writing model?

References

Biggs, J., & Watkins, D. A. (Eds.). (2001). *Teaching the Chinese lerner: Psychological and pedagogical perspective*. Hong Kong: Comparative Education Research Centre, University of Hong Kong.

Bloom, B. S., et al. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: Longman.

Byrant, S. L., John, L. T. S., Timmins, A. A., & Williams, J. A. H. (2002). *Formative assessment does enhancing learning: An annotated bibliography*. Hong Kong: Hong Kong Institute of Education.

Chen Kai. (2002). The guiding role of writing in extracurricular reading. *Foreign Language Teaching,* (2):29–33.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches*. Thousand Oaks: Sage.

Esmaeili, H. (2000). *The effects of content knowledge from reading on adult ESL students' written compositions in an English language test using reading and writing modules*. University of Toronto.

Freeman, D. L., & Long, M. H. (2000). *An introduction to second language acquisition Research*. Beijing: Foreign Language Teaching and Research Press.

Harmer, J. (2000). *How to teach English*. Beijing: Foreign Language Teaching and Research Press.

Janopoulous, M. (1986). The relationship of pleasure reading and second language writing proficiency. *TESOL Quarter, 20*(4), 763–768.

Johns, A. (1993). Reading and writing tasks in English for academic purposes classes: Products, processes, and resources. In J. Carson (Ed.), *Reading in the composition classroom: Second language perspectives* (pp. 274–289). Boston: Heinle & Heinle Publishers.

Krashen, S. J. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Joumal, 173*(4), 440–464.

Li, M. Y., & Lv, Q. (2002). Improving reading through writing. *Shandong Foreign Language Teaching Journal, 91*(6), 41–44.

Li, P. F., & Wang, L. X. (2011). Writing through reading and reading through writing. *Crazy English, 4*(3), 11–14.

Lorin, W. A., et al. (2001). *A taxonomy for learning, teaching and assessing – a revision of Bloom's taxonomy of educational objectives*. New York: Longman

Lorin, W. A., Krathwohl, D. R., et al. (2009). *A taxonomy for learning, teaching and assessing – a revision of Bloom's taxonomy of educational objectives*. Beijing: Foreign Language Teaching and Research Press.

MeMesser, S.D. (1997). *Evaluating ESL written summaries: An investigation of the ESL integrated summary profile (ISP) as a measure of the summary writing ability of ESL students.* The Florida State University.

Supervision Board for Foreign Language Teaching of Higher Educational Institutes (English Group). (2000). *The syllabus for undergraduate English majors in higher educational institutions*. Shanghai: Shanghai Foreign Language Education Press.

Tian, Y. (2003). *A study of practice and exploration of formative assessment in EFL middle teaching*. Doctoral thesis, Central China Normal University, Wuhan.

Wang, D. Q. (2007). *Authentic evaluation: From theory to practice*. Beijing: Foreign Language Teaching and Research Press.

Xu, H., & Gao, C. F. (2007). An empirical study on integrating reading with writing in EFL teaching. *Modern Foreign Languages, 30*(2), 184–190.

Yang, Y. L., & Dong, Y. Z. (2010). Writing through reading and reading through writing – a exploration of experiential teaching approach. *Foreign Languages in China, 7*(1), 13–27.

Yang, H., & Zhou, H. (2013). A study of reading-writing model based on Bloom's taxonomy of educational objectives. *Journal of Lishui University, 35*(1), 98–102.

Zhang, D. Y. (2005). English learning strategies and autonomous learning. *Foreign Language Education, 26*(1), 49–55.

Chapter 7
Learning by Assessing in an EFL Writing Class

Trevor A. Holster, William R. Pellowe, J. Lake, and Aaron Hahn

Abstract This study used many-faceted Rasch measurement to research peer assessment in EFL writing classes, following previous research which reported acceptance of the pedagogical benefits by students of attention paid to a rubric during peer assessment. Pre and post treatment writing was compared on two rubrics, one targeting specific instructional items, the second intended to measure general academic writing. Students used the instructional rubric to conduct peer assessment, but were not exposed to the secondary rubric. Substantively and statistically significant gains were observed on the instructional rubric but not on the secondary rubric, providing evidence of learning by assessing. Response patterns suggested holistic rating by peer raters, resulting in effective rank ordering of overall performances but an inability to provide formative feedback, supporting the view that the mechanism of learning was awareness arising from learning by assessing.

Keywords Peer assessment • Rasch model • Attention • Academic writing

7.1 Introduction

Standardized proficiency tests such as TOEFL (ETS 2008) aim to incorporate characteristics of language sampled from relevant real world tasks into test items (Chapelle 2008). Such tests focus on stable traits using multiple choice

T.A. Holster (✉) • J. Lake
Academic English Program, Fukuoka Women's University, Fukuoka, Japan
e-mail: trevholster@gmail.com

W.R. Pellowe
Department of Human Engineering, Kinki University, Higashiosaka, Japan

A. Hahn
Department of English, Fukuoka High School, Fukuoka, Japan

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_7,

tests, but many testing experts also support performance tests, integrating multiple skills into a single performance (see Brown et al. 2002; Hughes 2003; McNamara 1996, for example). Essay writing, for instance, requires integration of vocabulary, grammar, rhetorical structure, and content knowledge into a single performance. While performance tests can provide summative scores, they are also asserted to have formative benefits on study or teaching (Brown and Hudson 2002; Hughes 2003), but this requires that teachers and students become aware of weaknesses in their performances and are provided with opportunities to remedy them. Thus, validity arguments about formative tests require demonstration that the results are of sufficient quality to guide learning or instruction. One advantage of Rasch models (Bond and Fox 2007) over classical test theory (CTT) is the provision of fit statistics showing the consistency of students, items, and raters relative to the overall dataset, as demonstrated by Engelhard's (2009) investigation of students with disabilities. In the case of judged performance tests, many-faceted Rasch measurement (MFRM) (Linacre 1994) therefore provides invaluable diagnostic information about students, raters, or items in need of remediation.

In addition to feedback from teacher ratings, performance assessment used in conjunction with peer assessment (PA) and self assessment has potential formative benefits. Topping describes "learning by assessing" (LBA) as forcing increased time on task "interrogating the product or output, evaluating it in relation to intelligent questions at a macro and micro level" (1998, p. 254), following Graesser et al.'s (1995) investigation of discourse patterns in peer tutoring. Yarrow and Topping (2001) in their study of collaborative writing found that children made improvements in writing using a paired writing system that incorporated metacognitive prompting and scaffolding through the use of a framework in the form of a flowchart. They suggested that the flowchart acts as a metacognitive prompt, thus is both a tool for "procedural facilitation" during the writing process and for "product-oriented instruction" (p. 267) during editing and evaluating stages as well as for peer and self assessment. Similarly, Crinon and Marin researched French primary school children's writing, finding that children assigned to give feedback made larger improvements, "achieving greater overall coherency", than those who just received feedback, who "frequently make very specific local edits" (2010, p. 117). Likewise, Li et al. (2010) found that giving quality peer feedback was related to improved projects while receiving quality feedback did not improve the quality of final projects.

Further evidence of LBA was reported by Y. Cho and K. Cho (2011) who investigated English native speaker undergraduates writing technical reports, noting that "the effects of received peer comments were limited" (p. 639), but that writing improved more by giving comments than receiving comments. Further to this, K. Cho and MacArthur (2010) reported greater improvement for students who received feedback from multiple peers than from a single peer or single expert. The relative ineffectiveness of expert feedback suggests that it is engagement in the assessment process that provides the mechanism of learning rather than the product of assessment. Furthermore, K. Cho and MacArthur (2011)

compared outcomes for students who reviewed versus those who merely read the writing of peers and a control group, finding statistically and substantively better writing outcomes from the reviewing group, results that "clearly support the learning-by-reviewing hypothesis" (p. 77). K. Cho and Schunn (2010, pp. 209–210) noted that:

> The less obvious shift involves the change from students being the *receivers* of instructional explanations to students being the *generators* of instructional explanations. . . . In the role of reviewer, a student engages in reading, text analysis, and writing. . . . Coming to understand the criteria well enough to apply them to another student's paper provides students with the opportunity to improve their own writing and revision activities.

PA has attracted steady interest in second language acquisition. Mendonça and Johnson (1994) found that students not only made revisions due to peer comments but also self-noticed problems during peer review negotiations. They found that peer interactions forced students to be more active in their thinking about writing and this leads them to be able to use their knowledge in their revisions. Diab and Balaa (2011) used rubrics as instructional tools and for formative self assessment and peer assessment in EFL writing classes in Lebanon, finding statistically significant improvements in second-draft scores and strong endorsement from students of the value of the rubrics as learning tools. However, Min, studying Taiwanese university students, noted that although there are a large number of studies showing the benefits of peer response, "few studies have examined the extent to which peer feedback is incorporated into students' subsequent revisions" (2006, p. 119). Yang et al. (2006) compared teacher and peer feedback among Chinese EFL writing students, finding that teacher feedback was more likely to be incorporated and led to greater improvement, but also that it led to more superficial revisions. Peer feedback was found to be more likely to lead to meaning-change revision, a result attributed to negotiation of meaning during the peer interaction. Importantly, students in the peer feedback group had a much more positive view of its usefulness than students in the teacher feedback group.

Although a number of studies provide support for the formative benefits of peer assessment, the performance of student raters leads to doubts about its use for summative assessment. Roskams (1999) found support for LBA, but less support for assigning summative grades, while Tsui and Ng (2000) found that students preferred teacher responses over peer responses and incorporated more teacher feedback into revisions. They also found that students benefited more from reading the writing of peers than from peers' written comments, consistent with LBA. Mok (2011) found some acceptance of LBA among junior high school students, but serious concerns overall about the implementation of PA, while Cheng and Warren (2005) found students to be uncertain of their ability to rate peers and a tendency to rate holistically. Wong Mei Ha and Storey (2006) used journals to encourage reflection on self and peer editing, comparison with their own writing, and reflection on changes. Metacognitive awareness increased through reflection linking their declarative knowledge to their procedural knowledge. Saito and Fujita (2004)

compared peer and teacher ratings, finding a moderate to strong correlation, some support for LBA, but greater confidence in teacher ratings. Saito (2008) investigated the effect of rater training among Japanese university students, finding peer assessment to be effective overall, but a small effect from rater training and misfit patterns suggesting differential rubric interpretation between teachers and peer raters. Fukuzawa (2010) found acceptable fit for high school peer raters, but patterns suggesting a reluctance to use the lower categories on the rating scale, while Hirai et al. (2011) found undergraduate peer raters to be more lenient than teachers, as well as evidence of differential rating. Further concern over the consistency of peer raters was raised by Farrokhi et al.'s (2012) study of Iranian university students' English compositions, with student raters showing "a pattern of severity and lenience toward items that is opposite to that of teachers" (p. 93), and suggested the possibility that "they did not have a clear understanding of the assessment criteria."

Thus, although there is evidence supporting the effectiveness of LBA, serious doubts remain about students' understanding of the rubric, raising the question of the underlying mechanism by which LBA might aid second language acquisition. Schmidt (1990) argued that "noticing" driven by conscious attention is necessary for acquisition. Consistent with this, Schuchert (2004) argued that attention has a neurobiological basis, requiring alignment of five elements: an overall behavioral goal, a task-related goal, motor planning, stimulus qualities, and assessment of the influences of the four previous elements. This alignment produces the noticing required for both initial and advanced learning, so attention must be maintained in multiple sessions over extended timeframes for new knowledge to be consolidated as procedural knowledge. Although student raters may interpret the rubric differently to teachers, LBA may promote alignment of the elements of attention, leading to improved awareness of the rubric and noticing of the difference between students' own performances and target language features.

Thus, rather than the emphasis on the measurement of ability and consistency of test items typical of summative test analysis (Bachman 1990; Henning 1987), LBA's success rests on the quality of interaction of assessors with the rubric. Inconsistent assessors may have misunderstood the rubric or employed it idiosyncratically, casting doubt on their feedback to peers or ability to benefit from LBA. A validity argument for an assessment intended to generate LBA must therefore demonstrate that this interaction results in formative benefits independently of the quality of summative measurement. MFRM analysis provides an elegant solution to this, allowing peer raters' performances to be compared against those of teachers to ascertain whether the same trait is assessed by the different groups of raters, while also providing interval level measurement. The logit outputs provided by Rasch analysis provide convenient measures of effect size, allowing comparison between different studies on the basis of substantive meaningfulness, addressing Thompson's (1999) critique of misuse of measures of statistical significance, which are highly dependent on sample size.

7.2 The Study

7.2.1 Research Questions

RQ1: Do student performances improve after PA using the instructional rubric?
RQ2: Is peer feedback on specific rubric items comparable to teacher feedback?
RQ3: Are gains in the instructional rubric reflected in writing proficiency overall?

7.2.2 Background and Method

This study was conducted in writing classes in an Academic English Program (AEP) at a public women's university in Japan, but no familiarity with paragraph length organization could be assumed prior to this course. The participants ($N = 30$) in this convenience sampling were assigned to second semester classes taught by one of the authors, comprising 45 hours of instruction over 15 weeks. Although writing classes were conducted in two class groups of 15 students, all 30 students concurrently took a listening class together. The course book, *Ready to Write 3* (Blanchard and Root 2010), included brief grammar reviews which were used in class, but intensive grammatical instruction was not attempted due to time constraints and concerns over the sequencing and teachability of grammar features in general. Instruction targeted organizational features of writing considered to be learnable through explicit attention. Classes therefore focused on reviewing paragraph and essay organization and providing extensive writing practice on topics related to students' everyday experiences. Quite general topics were assigned, and students were encouraged to incorporate their personal experiences in order to maintain the relevance, novelty and coping potential argued by Schumann and Wood (2004) to underpin long-term motivation, while providing the alignment of the elements of attention described by Schuchert (2004).

Following reviews and practice of paragraph organization, the students planned and wrote three body paragraphs on the topic of "planning a trip that is educational, economical, and enjoyable". These were then combined into a complete essay following explanations of introductory and concluding paragraphs. A supplementary workshop on formatting conventions and the use of word processing software was conducted separately from the textbook curriculum, followed by peer review of the complete essays and submission of revised drafts for the PA session in the next class. Twenty-six students submitted Essay 1 in time, so these were randomly distributed to students for PA using the rubric shown in the Appendix.

PA generates very large numbers of responses, so data was collected using the peer assessment module for the open source MOARS audience response system (Pellowe 2002, 2010a, b). This system can output data ready for immediate MFRM analysis, making MFRM analysis practical within minutes of students completing their performances (Holster and Pellowe 2011). Paper rating sheets were used in

conjunction with MOARS, allowing items to be rated non-sequentially and providing a backup in case of technical issues with the online system. After approximately an hour of PA, students accessed the ratings for their own essay, presented in the form of bar graphs, and were asked to note strong and weak areas of their performances. Students were not provided the teacher's ratings separately from the PA results.

Division-classification essays were reviewed next, and then students were assigned the topic "Bad Habits" for Essay 2, required to plan three supporting paragraphs in class, and assigned a first draft for homework. In the next class, an introduction and conclusion were added and students were given a further week to produce a final draft. Twenty-four essays were submitted by the deadline; 23 students completed both Essay 1 and Essay 2.

7.2.3 Analysis and Results

Reliance of performance tests on human raters raises issues of rater performance. McNamara (1996) and Weigle (1994) provided seminal accounts of the use of MFRM to monitor rater performance and adjust for differences in severity. While teachers and students are implicitly familiar with two-faceted tests, where the probability of a successful response results from the interaction of student ability and item difficulty, judged performances introduce a third facet of measurement. The resulting probability of success is modeled as:

$$P = \exp(B - D - R)/(1 + \exp(B - D - R))$$

where P represents the probability of success, B represents person ability, D represents item difficulty, and R represents rater severity (Linacre 1994). Consistent with intuition, the odds of success increase with person ability, but decrease with item difficulty or rater severity. For the current study, a fourth facet, "Time", was included, on the hypothesis that student ability would increase as a result of the PA following Essay 1, and thus the probability of success would increase for Essay 2. This can be modeled as:

$$P = \exp(B - D - R - T)/(1 + \exp(B - D - R - T))$$

Three sets of teacher ratings were used for the initial analysis. Although the classroom teacher, T1, rated each essay as it was received, this resulted in multiple rating sessions over a period of several weeks, reflecting classroom reality, but raising concern over the consistency of the subsequent rating performance. T1 therefore rated all the essays again in a random order 1 week after the submission deadline for Essay 2. These ratings are indicated by T1A and T1B for the first and second ratings, respectively. For comparison, a second AEP teacher, T2, also rated all the essays following final submission.

MFRM analysis allowed measures of student ability, rater severity, proficiency gains across time, and item difficulty to be measured on a common log odds, or "logit", scale, representing equal interval measures, with items centered on 0.00 logits. Engelhard (2009) suggested 0.30 logits as a threshold for a substantive effect size, and all four facets showed substantively meaningful ranges. T1A was 0.68 logits more lenient than T2, an effect size translating into relative probabilities of success of 59 % versus 41 % for an item of average difficulty. The range of item difficulty, 2.31 logits, was extremely large. A student having a 50 % expectation of success with "Conclusion" (1.02 logits) would have a 91 % expectation of success on "Formatting" (−1.29 logits). Even the least proficient student (−0.66 logits) was substantively more able than "Formatting" is difficult, so this item provides little information about the ability of this sample of students. Finally, a provisional answer to RQ1 is possible by looking at the facet of "Time": student performances improved by approximately 1 logit following the PA session, a substantively very large gain.

However, definitive answers to the research questions assume acceptable functioning of all facets, so more detailed investigation is warranted. Fundamental to MFRM are assumptions of a unidimensional trait, so a preliminary question is whether teachers' ratings meet this requirement, given that these provide the benchmark against which PA ratings are compared. Item fit statistics provide an indication of whether the rubric describes a unidimensional trait, while rater fit statistics allow analysis of rater performance. Of particular concern was rubric Item 10, "Formatting", which addressed the use of word processing software rather than language proficiency. Rasch item analysis of the rubric was therefore conducted to determine whether psychometric evidence supported the content based argument concerning the dimensionality of the rubric. Table 7.1 shows the measurement report for items, ordered by model-data fit, shown by the infit and outfit mean-square (*MS*) statistics. "Formatting" is the most misfitting item, with infit and outfit statistics of 1.58 and 1.41 respectively. Given the questionable content validity of this item, this level of misfit supports removal of this item from the analysis.

Rater performance was investigated next, shown in Table 7.2. Raters T2 and T1B are slightly more consistent than expected, with mean-square statistics below 1.00, but T1A, with respective values of 1.17 and 1.14, is slightly misfitting. While this does not threaten overall measurement, the variation in performance between T1A and T1B shows the importance of multiple ratings for performance assessments. Subsequent analyses use only the ratings from T1B and T2.

Having demonstrated acceptable data-model fit for teacher ratings, the PA data was analyzed next. As peer feedback derived only from ratings of Essay 1, student ability measures for this essay were compared for teacher ratings and peer ratings, plotted in Fig. 7.1. Considerable agreement in rank ordering between teacher ratings and peer ratings is apparent, with a raw correlation of .87 indicating 75 % shared variance between the two sets of measures. With reliability coefficients for person measurement of .91 for peer ratings and .89 for teacher ratings, the disattenuated correlation rises to .97, indicating effectively interchangeable rank

Table 7.1 Item measurement report from ratings by teachers

Items		Score	*n*	*M*	Logit measure	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
10	Formatting	355	149	2.4	−1.29	0.14	1.58	1.41	.44
2	Introduction	224	150	1.5	0.73	0.12	1.21	1.20	.46
3	Conclusion	203	150	1.4	1.02	0.12	1.11	1.10	.40
1	Thesis stment	246	150	1.6	0.43	0.12	1.10	1.09	.57
6	Support	252	150	1.7	0.35	0.12	0.98	1.00	.26
7	Coherence	282	150	1.9	−0.08	0.12	0.90	0.90	.48
8	Cohesion	286	150	1.9	−0.14	0.12	0.89	0.89	.46
5	Unity	299	150	2.0	−0.33	0.12	0.83	0.82	.60
9	Relevance	269	150	1.8	0.11	0.12	0.80	0.80	.48
4	Organization	329	150	2.2	−0.80	0.13	0.73	0.74	.59
M ($n = 10$)		274.5	149.9	1.8	0.00	0.12	1.01	1.00	.47
SD (Pop)		43.9	0.3	0.3	0.66	0.01	0.24	0.20	.10
SD (Sample)		46.2	0.3	0.3	0.69	0.01	0.25	0.21	.10

Model(Pop): RMSE .12 Adj (True) SD.65 Separation 5.30 Strata 7.40 Reliability .97
Model(Samp): RMSE .12 Adj (True) SD.68 Separation 5.60 Strata 7.80 Reliability .97
Model fixed (all same) chi-square: 271.7 *df*: 9 significance (probability): .00

Table 7.2 Teacher rater's measurement report

Raters	Score	*n*	*M*	Logit meas	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
T1A	881	450	2.0	−0.43	0.07	1.17	1.14	.59
T2	773	450	1.7	0.12	0.07	0.94	0.94	.43
T1B	736	450	1.6	0.31	0.07	0.89	0.90	.57
M ($n = 3$)	796.7	450.0	1.8	0.00	0.07	1.00	0.99	.53
SD (Pop)	61.5	0.0	0.1	0.31	0.00	0.12	0.11	.07
SD (Sample)	75.3	0.0	0.2	0.38	0.00	0.15	0.13	.09

Model(Pop): RMSE .07 Adj (True) SD .31 Separation 4.29 Strata 6.05 Reliability .95
Model(Samp): RMSE .07 Adj (True) SD .38 Separation 5.30 Strata 7.40 Reliability .97
Model fixed (all same) chi-square: 57.2 df: 2 significance (probability): .00
Inter-Rater Exact Agreements: 595 = 44.1% Expected: 552.4 = 40.9%

ordering within the limits of measurement error. However, it is also apparent from Fig. 7.1 that peer raters were much more lenient than teachers, with mean logit measures of 1.79 for peer ratings versus −0.04 for teachers, corresponding to mean raw ratings of 2.3 versus 1.5 on the rating scale of 0–3.

In contrast, although the teachers and peer raters returned very similar rank ordering of person measures, this did not hold for the ranking of item difficulty measures, as shown in Fig. 7.2. While the range of item difficulty from teacher ratings was 2.12 logits, the range from peer ratings was only 0.83 logits, raising doubts about peer raters' interpretation of the rubric. The respective mean rating for

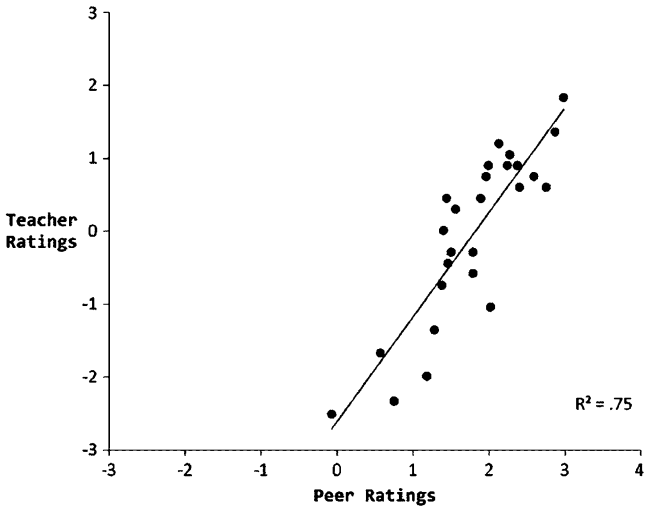


Fig. 7.1 Person ability measures from teacher ratings and peer ratings. Ratings of writing ability made by teachers and peers are mapped, showing a strong linear trend-line with shared variance of 75 %

teachers and peer raters were 1.5 and 2.4, with standard deviations of 0.2 and 0.1, and ranges of 0.8 and 0.3. Peer raters avoided the lower categories on the rating scale, while teachers were more likely to utilize the full range of the scale. These results are consistent with holistic ratings by peers, resulting in discrimination of good performances from poor, but not between the items on the rubric. Thus, RQ2 is answered: peer feedback on specific rubric items is not comparable to teacher feedback. Given that teacher feedback was not provided in this case, the inability of peer raters to provide diagnostic feedback raises the question of the source of the large gain between Essay 1 and Essay 2.

Figure 7.3 plots the interaction between items and time from teacher ratings, with all rubric items receiving higher mean ratings for Essay 2, reflected in the lower measures of difficulty. This further confirms RQ1: student performances following PA improved substantively. Only one item, "Thesis stment", with a gain of 2.18 logits, showed substantively larger improvement than the mean of 1.04 logits, but comparison with Fig. 7.2 shows that peer raters rated this as relatively easy, unlike teachers who rated it as difficult. Peer feedback cannot therefore have signaled to students that this item needed remediation, evidence against peer feedback as a major mechanism of improvement.

Given that Essay 1 was many students' first attempt at writing essay length compositions, the question arises whether the substantive overall gains arose from practice rather than LBA. Therefore a secondary rubric was developed independently to measure general writing proficiency. A writing instructor, R1, with a Masters degree in writing instruction and experience teaching academic writing to both North American university undergraduates and L2 learners in Japan was shown the submissions for

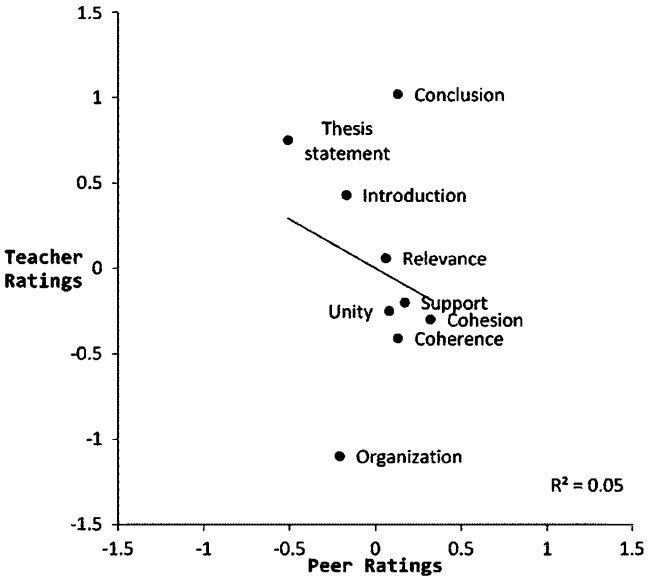


Fig. 7.2 Item difficulty measures from teacher ratings and peer ratings. The difficulty of rubric items estimated from teacher ratings and peer ratings are mapped, showing no correlation between the two sets of measures

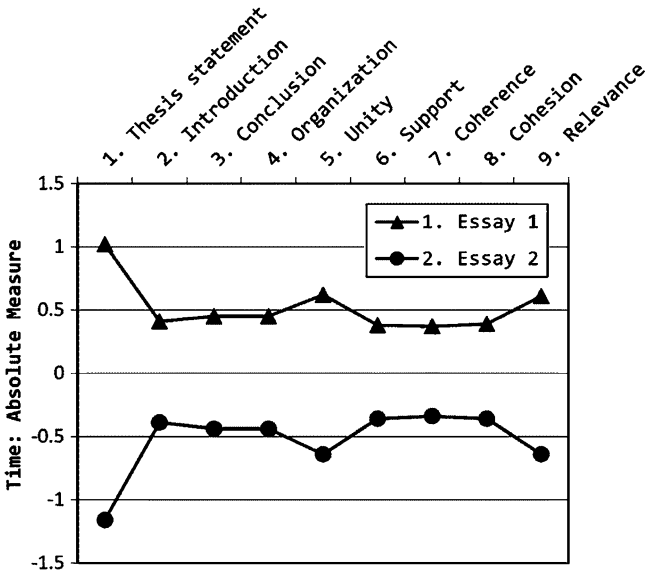


Fig. 7.3 Change in item difficulty by time for instructional rubric. The difficulty of rubric items is compared for Essay 1 and Essay 2. Ratings were substantively higher on all items in Essay 2, evidence of learning related to the rubric items

Table 7.3 Secondary rubric raters' measurement report

Raters	Score	*n*	*M*	Logit meas	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
R1	670	195	3.4	0.01	0.07	1.16	1.16	.62
R2	691	200	3.5	−0.01	0.07	0.83	0.83	.40
M ($n = 2$)	680.5	197.5	3.4	0.00	0.07	0.99	0.99	.51
SD (Pop)	10.5	2.5	0.0	0.01	0.00	0.17	0.17	.11
SD (Sample)	14.8	3.5	0.0	0.02	0.00	0.24	0.24	.15

Model (Pop): RMSE .07 Adj (True) SD .00 Separation .00 Strata .33 Reliability .00
Model (Samp): RMSE .07 Adj (True) SD .00 Separation .00 Strata .33 Reliability .00
Model fixed (all same) chi-square: .1 *df*: 1 significance (probability): .82
Inter-rater agreement opportunities: 195 Exact agreements: 52 = 26.7 % Exp: 53.8 = 27.6 %

Essay 1, but told only that students were taking academic writing classes and that they were asked to write an essay on the topic of "Planning a Trip", with introductory and concluding paragraphs and a minimum of three supporting paragraphs. The resulting rubric was based on experience with the writing section of the GRE (ETS 2012) and comprised four operational items, "Grammar", "Organization/Structure/Length", "Vocabulary/Register/Tone", and "Content/Logic/Context", rated on a scale from 1 to 6. A second rater, R2, was used to provide inter-rater comparison, this rater having an undergraduate degree in literature and a Masters degree in applied linguistics, with over two decades experience teaching L2 learners of English in North America and Japan.

Table 7.3 shows the rater measurement report for the secondary ratings. An inconsequential difference of 0.02 logits was found in severity. R1 was slightly less consistent than expected, with infit and outfit mean square statistics of 1.16 indicating 16 % more randomness than modeled, while R2 was correspondingly overfitting, with infit and outfit statistics of 0.83, levels not threatening to effective measurement. Table 7.4 shows the measurement report for items. The mean of both infit and outfit was 0.99, both having extremely low standard deviations of 0.09, while the most misfitting item was "Grammar", with infit and outfit mean square statistics of 1.13. Reliability of person measurement was .78, so these items are functioning acceptably and able to separate low ability persons from high.

Analysis of the secondary ratings found Essay 1 to be 0.08 logits more difficult than Essay 2, a gain that was neither statistically nor substantively significant. The gains measured in the PA rubric were not replicated, consistent with gains resulting from explicit awareness of the rubric rather than general proficiency. RQ3 is thus answered: gains in the instructional rubric were not replicated in the secondary rubric. However, Fig. 7.4 shows the time versus item interaction, with "Grammar" given lower ratings and "Organization/Structure/Length" given higher ratings in Essay 2. Drawing firm conclusions about this from such a small pilot dataset is inadvisable, but it is notable that "Organization/Structure/Length" is similar to PA rubric items, while the others are not, consistent with PA leading to LBA.

Table 7.4 Secondary rubric items' measurement report

Items	Score	*n*	*M*	Logit meas	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
2 Grammar	339	99	3.4	0.02	0.10	1.13	1.13	.51
3 Org/struc/length	354	98	3.6	−0.16	0.10	0.99	0.98	.51
4 Vocab/reg/tone	359	99	3.6	−0.17	0.10	0.88	0.90	.47
5 Cont/logic/context	309	99	3.1	0.32	0.10	0.96	0.96	.51
M ($n = 4$)	340.3	98.8	3.4	0.00	0.10	0.99	0.99	.50
SD (Population)	19.5	0.4	0.2	0.20	0.00	0.09	0.09	.02
SD (Sample)	22.5	0.5	0.2	0.23	0.00	0.10	0.10	.02

Model (Pop): RMSE .10 Adj (True)SD .17 Separation 1.72 Strata 2.63 Reliability .75
Model (Samp): RMSE .10 Adj (True)SD .21 Separation 2.07 Strata 3.09 Reliability .81
Model, Fixed (all same) chi-square: 15.8 df: 3 significance (probability): .00

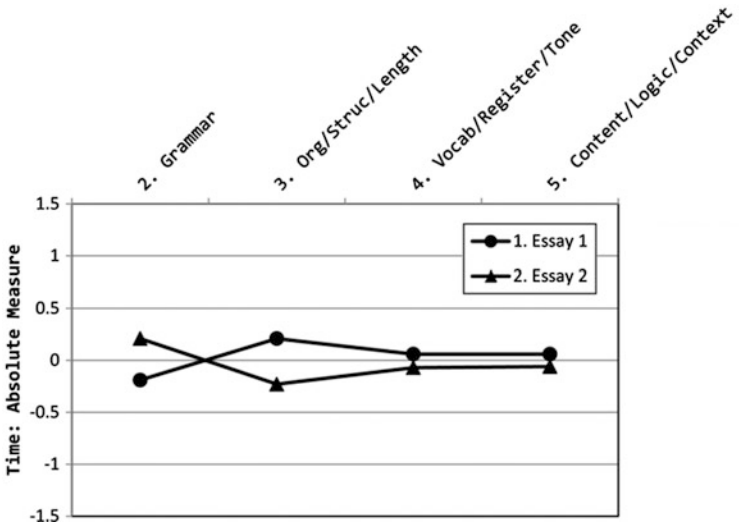


Fig. 7.4 Change in item difficulty by time for secondary rubric. The difficulty of rubric items is compared for Essay 1 and Essay 2, with a substantive gain on Item 3 offset by a substantive loss on Item 2

7.3 Discussion and Implications

The major research question, RQ1, concerned the effectiveness of PA leading to improved performance on Essay 2. The results supported this, a substantively large effect size occurring between Essay 1 and Essay 2 on the PA rubric but not on the secondary rubric, consistent with LBA. As teacher feedback was not provided on

the rubric items and teachers and peer raters employed the rubric differently, peer feedback could not have identified rubric items in need of remediation, leaving attention to the rubric during the rating sessions as the most plausible source of learning. Although the research design did not control for the difficulty of the essay topics, raising the concern that the higher ratings for Essay 2 on the PA rubric may have resulted from the second topic being easier than the first, this pattern was not seen in the results from the secondary rubric. This supports LBA as a powerful mechanism of learning through drawing attention to key features of performances.

These results support the validity of peer assessment as a classroom instructional task while holding potential benefits for motivation because it was interaction with samples of student language that resulted in LBA. This expands the input available to learners and addresses the argument for a balance between familiarity and novelty made by Schumann and Wood (2004) by providing input on topics that are relevant and interesting while promoting the alignment of the elements of attention described by Schuchert (2004).

However, the results of this pilot study were limited by sampling constraints and the limited timeframe. The classroom teacher's impressionistic feeling was that these students had very high intrinsic motivation and were not representative of average Japanese university students. Although large gains were observed on the PA rubric between Essay 1 and Essay 2, a ceiling effect may occur if further rounds of writing and PA were administered, while it's plausible that larger gains on the secondary rubric would be observed over a longer timeframe. Furthermore, the vagueness of the essay prompts, intended to provide students with opportunities to write about familiar topics and experiences, were not well suited to the secondary rubric, based on the expectations of L1 academic writing. Addressing these issues was beyond the scope of this pilot study, but highlight the need for a larger scale quasi-experimental study to confirm these findings and provide evidence of wider generalizability.

Appendix

Essay Rating Instructions and Rubric

Essay Revision

Read other students' essays. Rate each essay from "A" to "D" on the following points by marking the bubbles on the grading sheet.

他の学生の発表を見て評価をします。以下の評価基準を参考にして、評価シートのA~Dをりつぶして下さい。

"A" = Excellent performance.

(素晴らしい。)

"B" = Good performance, but could be improved.

(良いが、改善出来る部分もある。)

"C" = Weak performance, should be improved.

(良いとは言えない。改善した方が良い。)
"D" = Very weak performance, must be improved.
(良くない。改善すべき。)

1. **Thesis stment**: How well does the introduction identify the focus of the essay using a thesis stment?
2. **Introduction**: How well does the introduction preview the main points of the essay?
3. **Conclusion**: How well does the conclusion summarize the main points of the essay?
4. **Organization**: Are the supporting paragraphs in a logical order?
5. **Unity**: Does each supporting paragraph have a clear topic sentence and focus?
6. **Support**: Do the supporting paragraphs support the essay focus with specific details?
7. **Coherence**: Are the supporting sentences in each paragraph organized in a logical way?
8. **Cohesion**: Did the writer use transition words to guide the reader from one idea to the next?
9. **Relevance**: Are all the supporting sentences relevant?
10. **Formatting**: Is the essay formatted correctly?

References

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

Blanchard, K., & Root, C. (2010). *Ready to write 3: From paragraph to essay* (3rd ed.). White Plains: Pearson Longman.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum Associates.

Brown, J. D., & Hudson, T. D. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments*. Honolulu: University of Hawaii.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York: Routledge.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22* (1), 93–121. doi:10.1191/0265532205lt298oa.

Cho, Y., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39* (5), 629–643. doi:10.1007/s11251-010-9146-1.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–338.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103* (1), 73–84. doi:10.1037/a0021950.

Cho, K., & Schunn, C. D. (2010). Developing writing skills through students giving instructional explanations. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines*. New York: Springer.

Crinon, J., & Marin, B. (2010). The role of peer feedback in learning to write explanatory texts: Why the tutors learn the most. *Language Awareness, 19*(2), 111–128. doi:10.1080/09658411003746604.

Diab, R., & Balaa, L. (2011). Developing detailed rubrics for assessing critique writing: Impact on EFL university students' performance and attitudes. *TESOL Journal, 2*(1), 52–72. doi:10.5054/tj.2011.244132.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585–602. doi:10.1177/0013164408323240.

ETS. (2008). The TOEFL® Test – Test of English as a Foreign Language™. Retrieved from http://tinyurl.com/zocgc. Accessed 28 Mar 2008.

ETS. (2012). About the GRE® revised General Test. Retrieved from http://www.ets.org/gre/revised_general/about. Accessed 19 Jan 2012.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79–101.

Fukuzawa, M. (2010). Validity of peer assessment of speech performance. *Annual Review of English Language Education in Japan, 21*, 181–190.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495–522. doi:10.1002/acp.2350090604.

Henning, G. (1987). *A guide to language testing*. Boston: Heinle & Heinle.

Hirai, A., Ito, N., & O'ki, T. (2011). Applicability of peer assessment for classroom oral performance. *JLTA Journal, 14*, 41–59.

Holster, T. A., & Pellowe, W. R. (2011). *Using a mobile audience response system for classroom peer assessment*. Paper presented at the JALT CALL 2011 conference, Kurume University, Kurume.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology, 41*(3), 525–536. doi:10.1111/j.1467-8535.2009.00968.x.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Pearson Education.

Mendonça, C. O., & Johnson, K. E. (1994). Peer review negotiations: Revision activities in ESL writing instruction. *TESOL Quarterly, 28*(4), 745–769.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118–141. doi:10.1016/j.jslw.2006.01.003.

Mok, J. (2011). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal, 65*(3), 230–239. doi:10.1093/elt/ccq062.

Pellowe, W. R. (2002). *Keitai-assisted language learning (KALL)*. Paper presented at the 28th JALT international conference, Granship Conference Center, Shizuoka.

Pellowe, W. R. (2010a). MOARS (Version 0.8.3) [Audience response system]. Retrieved from http://moars.com

Pellowe, W. R. (2010b). *Quiz and survey system for mobile devices*. Paper presented at the 36th JALT international conference, WINC, Aichi.

Roskams, T. (1999). Chinese EFL students' attitudes to peer feedback and peer assessment in an extended pairwork setting. *RELC Journal, 30*(1), 79–123. doi:10.1177/003368829903000105.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553–581. doi:10.1177/0265532208094276.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31–54. doi:10.1191/1362168804lr133oa.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. doi:10.1093/applin/11.2.129.

Schuchert, S. A. (2004). The neurobiology of attention. In J. H. Schumann, S. E. Crowell, N. E. Jones, N. Lee, S. A. Schuchert, & L. A. Wood (Eds.), *The neurobiology of learning* (pp. 143–174). Mahway: Lawrence Erlbaum Associates.

Schumann, J. H., & Wood, L. A. (2004). The neurobiology of motivation. In J. H. Schumann, S. E. Crowell, N. E. Jones, N. Lee, S. A. Schuchert, & L. A. Wood (Eds.), *The neurobiology of learning* (pp. 23–42). London: Lawrence Erlbaum Associates.

Thompson, B. (1999). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191–196. doi:10.1177/095935439992007.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. doi:10.3102/00346543068003249.

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147–170.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223. doi:10.1177/026553229401100206.

Wong Mei Ha, H., & Storey, P. (2006). Knowing and doing in the ESL writing class. *Language Awareness, 15*(4), 283–300.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.

Yarrow, F., & Topping, K. J. (2001). Collaborative writing: The effects of metacognitive prompting and structured peer interaction. *The British Journal of Educational Psychology, 71*, 261–282.

Chapter 8
Construction and Evaluation of an Item Bank for an Introductory Statistics Class: A Pilot Study

Sieh-Hwa Lin, Pei-Jung Hsieh, and Li-Chuan Wu

Abstract The purpose of the present study is to construct and evaluate the applicability of an item bank for an Introductory Statistics class. The participants of the study were 54 college students enrolled in Statistics in Psychology and Education in the spring and fall semesters of 2010. To establish the test bank, the authors first adopted and re-wrote the items from the midterm and final tests from the previous school year. Students practice the test in advance of the class on the instructional platform, Moodle. A total of 15 units were prepared with 45 items. The results revealed that (1) the point-biserial correlation of 34 items (75.6 %) reached .25, meaning the test items constructed for this study had enough discriminatory power; (2) 80 % of the Rasch item difficulty values ranging from −1 to 1, indicating an appropriate difficulty of the test bank, which not only met the students level, but was also appropriate to help students with preparation for the unit. Students' achievement performance on the course correlated positively with the number of times they took the preparation tests, suggesting that the implementation of the lesson preparation activity enhanced the learning effectiveness of statistics.

Keywords Introductory statistics • Item bank • Rasch model

S.-H. Lin • L.-C. Wu
Department of Educational Psychology and Counseling, National Taiwan Normal University, Taipei, Taiwan

P.-J. Hsieh (✉)
Research Center for Testing and Assessment, National Academy for Educational Research, No. 2 Sanshu Rd., Sanxia District, New Taipei City 23703, Taiwan
e-mail: pjh@mail.naer.edu.tw

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_8,

8.1 Introduction

In recent years, all kinds of wireless facilities have become more and more widespread with the flourishing development of wireless technology. No matter the public transportation systems or personal mobile communication devices, all attempts to provide services to users to send, transmit, or receive symbols, signals, texts, images, sounds, or other kinds of messages with wireless broadband technology. In order to encourage students to use information technology to enhance their learning and living abilities, the Ministry of Education in Taiwan provides subsidies to schools to install wireless hotspots so as to improve the fundamental application environment. With such support, educators are exploring the possibilities to apply mobile learning, or m-learning, in the field of education.

After reviewing relevant studies in educational psychology, Calfee (2006) pointed out that teaching in the twenty-first century should solve crucial education problems which include shortening achievement gaps between different groups, providing effective and efficient educational activities, studying the correlation between neurophysiology and learning in depth, and carrying out studies regarding school organizations and leadership. Calfee thinks that providing effective and efficient teaching activities by using m-learning is one of the educational emphases.

Statistics is a field of science which is characterized by the accumulated knowledge. Students often feel a high cognition burden during learning process. Previous studies point out that students can acquire immediate feedback during the concept acquisition stage if they are provided with timely questioning and evaluation in class, and thereby understand his or her learning performance. Using mobile devices as evaluation tools can increase the quality of immediate feedback to an even greater extent (Jones et al. 2002). As a result, using mobile learning in statistics courses may not only help students to understand their own learning status immediately, but also help teachers to improve on things in a timely manner that students might have more difficulty in framing concepts for.

However, Wen et al. (2008) analyzed 42 research projects which were titled "mobile learning" and approved by National Science Council of the Executive Yuan from 2003 to 2007. They discovered that 12 (28.57 %) of them focused on "the installation of mobile learning platforms", and that 7 (16.66 %) of them centered on "strategies, theories, design, evaluation, and management of mobile learning". Consequently, one can conclude that there is not much domestic research on the application of m-learning in classroom assessment, let alone its application in statistics courses.

To achieve the long-term objective of mobile evaluation of introductory statistics, this project established a test bank required for the study in a preliminary stage for the classroom assessment of m-learning in the future. As a result, the major purposes of this study are:

- To construct online test items for introductory statistics.
- To evaluate the applicability of online test items for introductory statistics.

8.2 Method

8.2.1 Participant

The participants in this study are 54 students enrolled in Psychology and Education during the spring and fall semesters of 2010. Female students are the majority, and most of them major in Education Psychology and Counseling.

8.2.2 Measurement

Test items used in this study were revised from the two midterm exams and the two final exams from the 2009 school year, 35 items for the midterms and 50 for the finals. Furthermore, the responses the students from that school year for the four tests were taken as the bases. The results of the question analyses of the four tests were taken as references to select appropriate items from each unit and rewrite them.

8.2.3 Procedure

The textbook is *Understanding Statistics in the Behavioral Sciences* of Pagano (2010), and the scope is from chapter 3 to chapter 17. The test bank is based on the progress of the teaching plan. Three items from each unit will be provided on the Moodle teaching platform 1 week prior to the class for the unit, so students can answer them before the class. Tests will be closed when the teaching of that unit is finished. The items are in random order, and so are the options for every unit.

8.2.4 Analysis

The Rasch model is used to estimate the test parameters and ability parameters of students. Appropriate items will be saved in the test bank after analysis. The results of the tests can be used to estimate not only the parameters but also the students' weak points in learning in order to improve at an early stage. The analysis software is Winsteps 3.70.

8.3 Results and Discussion

Table 8.1 shows that the percentage of students to take the online test was above 80 % for every unit but one. Only one test was less than 80 %, at 72.22 %. It is obvious that the majority of students continued to participate in the online test.

Table 8.1 Statistics of number of people participating in the tests ($N = 54$)

Test unit	n (%)
Ch 03 Frequency distributions	46 (85.19 %)
Ch 04 Measures of central tendency and variability	52 (96.30 %)
Ch05 The normal curve and standard scores	52 (96.30 %)
Ch06 Correlation	50 (92.59 %)
Ch07 Linear regression	51 (94.44 %)
Ch08 Random sampling and probability	44 (81.48 %)
Ch09 Binomial distribution	51 (94.44 %)
Ch10 Introduction to hypothesis testing using sign test	51 (94.44 %)
Ch11 Power	39 (72.22 %)
Ch12 Sampling Distributions, Sampling Distribution of the Mean, the Normal Deviate (z) Test	45 (83.33 %)
Ch13 Student's t test for single samples	50 (92.59 %)
Ch14 Student's t test for correlated and independent groups	50 (92.59 %)
Ch15 Introduction to the analysis of variance	47 (87.04 %)
Ch16 Introduction to two-way analysis of variance	52 (96.30 %)
Ch17 Chi-squares and other nonparametric tests	49 (90.74 %)

The items are classified as statistical concepts and calculations according to their attributes. Items that do not require any formula to select the answers are considered statistical concepts. If any formula is required to find the answers, items are classified as calculations. There are 25 items (55.56 %) which are statistical concepts among the 45 items; 20 items (44.44 %) are calculation items.

Table 8.2 is the outcome of analyses with regard to question quality. The level of difficulty, δ value, of 80 % of items lies in between −1 and 1, meaning the level of difficulty for the online tests is appropriate and is in accordance with the level of the students. There are ten items whose point-biserial correlation is less than .25. Six of them are statistical concepts, and five are calculations. Also, the δ values of three of them are greater than one and are items with a higher level of difficulty. The three items are derived from chapter 10, "Introduction to hypothesis testing using sign test," chapter 11, "Power," and chapter 16, "Introduction to two-way analysis of variance," respectively. In chapter 10 we begin to enter the field of research hypothesis. Students have to learn to adjust their logical thinking, so it is easier to understand. Chapter 11, "Power," is about very abstract concepts. Considering the level of difficulty of all units for the entire year, chapter 16, "Introduction to two-way analysis

Table 8.2 Item difficulty and point-biserial correlation of each item

Item	δ	Point-biserial correlation	Item attribute
Ch 03 Frequency distributions			
Ch03-01	.87	.21	Statistical concepts
Ch03-02	−.43	.37	Statistical concepts
Ch03-03	−.43	.08	Calculation
Ch 04 Measures of central tendency and variability			
Ch04-01	.66	.67	Statistical concepts
Ch04-02	.17	.50	Calculation
Ch04-03	−.82	.43	Statistical concepts
Ch05 The normal curve and standard scores			
Ch05-01	.13	.48	Statistical concepts
Ch05-02	−1.63	.46	Calculation
Ch05-03	1.50	.51	Statistical concepts
Ch06 Correlation			
Ch06-01	−.93	.41	Statistical concepts
Ch06-02	1.45	.24	Statistical concepts
Ch06-03	−.53	.39	Calculation
Ch07 Linear regression			
Ch07-01	−.91	.46	Statistical concepts
Ch07-02	1.02	.33	Calculation
Ch07-03	−.11	.31	Calculation
Ch08 Random sampling and probability			
Ch08-01	.00	.18	Calculation
Ch08-02	.00	.49	Statistical concepts
Ch08-03	.00	.49	Calculation
Ch09 Binomial distribution			
Ch09-01	−.93	.67	Calculation
Ch09-02	−3.16	.68	Calculation
Ch09-03	4.09	.51	Calculation
Ch10 Introduction to hypothesis testing using sign test			
Ch10-01	−.28	.31	Statistical concepts
Ch10-02	−1.22	.23	Statistical concepts
Ch10-03	1.49	.06	Calculation
Ch11 Power			
Ch11-01	−.11	.46	Statistical concepts
Ch11-02	−1.66	.33	Statistical concepts
Ch11-03	1.77	.12	Statistical concepts
Ch12 Sampling Distributions, Sampling Distribution of the Mean, the Normal Deviate (z) Test			
Ch12-01	−.09	.32	Statistical concepts
Ch12-02	−.64	.64	Statistical concepts
Ch12-03	.73	.55	Statistical concepts
Ch13 Student's t test for single samples			
Ch13-01	.82	.32	Statistical concepts
Ch13-02	−.22	.06	Statistical concepts
Ch13-03	−.60	.27	Statistical concepts

(continued)

Table 8.2 (continued)

Item	δ	Point-biserial correlation	Item attribute
Ch14 Student's t test for correlated and independent groups			
Ch14-01	−.11	.44	Statistical concepts
Ch14-02	.06	.10	Statistical concepts
Ch14-03	.06	.40	Calculation
Ch15 Introduction to the analysis of variance			
Ch15-01	.53	.29	Calculation
Ch15-02	−.07	.37	Calculation
Ch15-03	−.46	.29	Calculation
Ch16 Introduction to two-way analysis of variance			
Ch16-01	−.54	.37	Statistical concepts
Ch16-02	−.54	.23	Calculation
Ch16-03	1.08	.07	Calculation
Ch17 Chi-squares and other nonparametric tests			
Ch17-01	.25	.37	Statistical concepts
Ch17-02	.00	.56	Calculation
Ch17-03	−.25	.68	Calculation

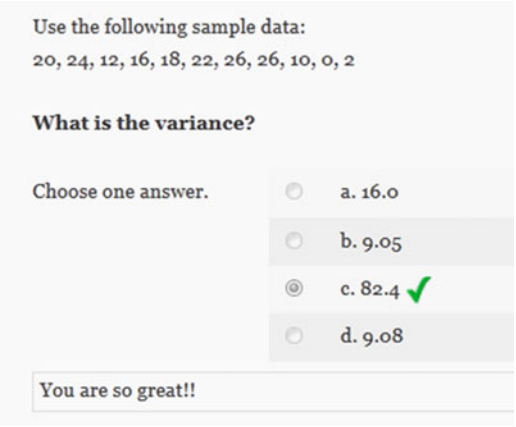


Fig. 8.1 An example of feedback for answering correct

of variance" is also a more difficult part. As the statistics class progresses and the test is designed, these few units should be explained more carefully.

In addition to constructing the tests, this study also focuses on discussing the advantages of online tests, which is the "feedback" mechanism. Through the feedback function of the Moodle platform, students can receive corresponding feedback messages no matter whether the answer is correct or wrong. If the answer is correct, positive and encouraging texts are shown (Fig. 8.1); if the answer is wrong, sources in the text book will be pointed out and an explanation will be presented so that students can compare it with their blind points (Fig. 8.2).

Using the number of preview tests taken in the first and second semesters in 2010 school year and scores in the midterms and finals to carry out a relevancy analysis (Tables 8.3 and 8.4), it is discovered that the more preview tests taken, the score in

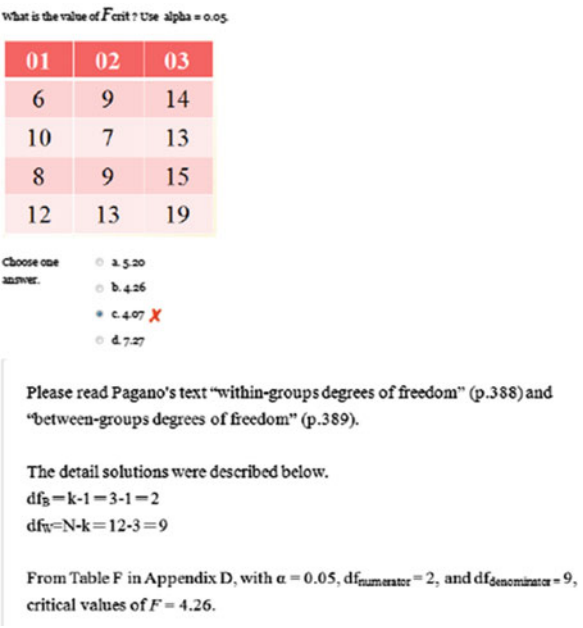


Fig. 8.2 An example of feedback for answering incorrect

Table 8.3 Correlation between login times and statistical achievement in the fall semester of 2010

	(1)	(2)	(3)
Number of preview tests taken prior to midterm	–		
Midterm score	.30*	–	
Preview tests taken in the entire semester	.86**	.41**	–
Final score	.40**	.73**	.60**

Note: *$p < .05$; **$p < .01$

Table 8.4 Correlation between login times and statistical achievement in the spring semester of 2010

	(1)	(2)	(3)
Number of preview tests taken prior to midterm	–		
Midterm score	.55*	–	
Preview tests taken in the entire semester	.93*	.60*	–
Final score	.48*	.72*	.61*

Note: *$p < .01$

statistics is higher, and the correlation between the two lies between .30 and .61, which reached relevancy above the medium degree. Therefore, it is obvious that the implementation of preview tests has positive effects on enhancing the learning effectiveness of students in statistics classes.

On the other hand, "number of preview tests taken prior to midterm" and "number of preview tests taken in the entire semester" are highly correlated. The correlation of the two in the fall semester of 2010 is $r = .86$, $p < .01$. The correlation of the two in the spring semester of 2010 is $r = .93$, $p < .01$. This shows that participating in online tests is a learning behavior that can be cultivated. As long as students are encouraged at the beginning of the course to take part in online tests, it is possible to form such a long-term habit.

In the future, online tests can be integrated into one of the teaching activities in the introductory statistics class to encourage students to go online and participate in the tests. Besides increasing the self-learning ability, it might also help improve learning performance.

References

Calfee, R. (2006). Educational psychology in the 21st century. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 29–42). Mahwah: Lawrence Erlbaum Associates.

Jones, C. G., Johnson, D. W., & Cold, S. J. (2002). M-education: Mobile computing enters the classroom. *Issues in Information Systems, 3*, 309–315.

Pagano, R. R. (2010). *Understanding statistics in the behavioral sciences* (9th ed.). Belmont: Wadsworth.

Wen, J.-R., Sun, S.-Y., Huang, C.-H., & Kuo, S.-H. (2008, May). *Mobile learning research and trends: The Evaluation of National Science Council projects during 2003–2007*. Paper presented at the third mobile and ubiquitous technologies enhanced learning conference, National University of Tainan, Taiwan.

Chapter 9
The Impact of Unobserved Extreme Categories on Item and Person Estimates – A Simulation Study

Edward Feng Li

Abstract For any polytomous items, it sometimes occurs that an extreme category, which is logically possible, is not observed in a particular sample. For example, in education when the performance tasks by the students from different year levels are judged by the same set of criteria, it is likely that none of the lower year level students would achieve the highest marks in some criteria. In health, it may happen when a group of generally healthy participants are measured by an instrument designed to detect some particular symptoms. This paper uses a simulation study to investigate the impact of unobserved extreme categories on item and person estimates. Based on the polytomous Rasch model, the Partial Credit Model (Masters, Psychometrika 47:149–174, 1982), data were simulated for 1,000 persons, N(0,1), and ten polytomous items with five categories under two scenarios, one with unobserved extreme high categories and the other with unobserved extreme low categories. The generated data sets were analysed with the RUMM2030 software (Andrich et al. 2009). The results show that unobserved extreme high categories in the data tend to lead to overestimated person means and unobserved extreme low categories tend to lead to underestimated person means. Both scenarios resulted in underestimated item standard deviations. The results suggest that collapsing unobserved extreme categories improves person and item estimation accuracy, especially when a large proportion of items have an unobserved extreme category. These results have implications for designing and measuring performance tasks that need to be carried out across a wide spectrum of ability groups. It may also affect the common-item equating procedure for polytomous items where the threshold values of common items are used.

Keywords Rasch model • Partial Credit Model • Polytomous items • Unobserved extreme categories • Null categories • Collapsing categories • Estimation accuracy • Bias • RMSE • Simulation

E.F. Li (✉)
University of New South Wales, Kensington, NSW, Australia
e-mail: edward.f.li@unsw.edu.au

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_9,

9.1 Introduction

As a modern psychometric approach, the Rasch measurement model has been adopted in many areas, such as education and medical and health care assessment, to evaluate outcomes (see, for example, Andrich 1988; Bond and Fox 2001; Hobart and Cano 2009; Linacre et al. 1994; Pallant and Tennant 2007; Rasch 1960; Tennant and Conaghan 2007; Tennant et al. 2004; Wright and Linacre 1989). The wide variety of applications of the Rasch model may be attributed to its desired property of invariance, because this principle of invariance of comparisons is a feature of fundamental measurement (Andrich 1989; Thurstone 1928; Wright and Stone 1999). It states that, given the same frame of reference, the comparison between any two items is independent of which persons are chosen and the comparison between any two persons is independent of which items are used (Rasch 1961, cited in Andrich 1989). Moreover, it is argued that the Rasch model has revolutionised the field by redefining the data-model relationship in measurement, the Rasch paradigm (Andrich 2004). In the traditional paradigm, statistical models are developed to characterise the data. As generally there is no a priori restriction on what class of models to use, how well models account for the data becomes a main criterion to select a model. However, in the Rasch paradigm, what model might be used subscribes to certain criteria of measurement required by scientists (Duncan 1984; Kuhn 1961/1977). The basic Rasch model is applicable to dichotomous items which can be marked as right (1) or wrong (0). It can be expressed in the form

$$\Pr\{X_{ni} = x\} = [\exp(x(\beta_n - \delta_i))]/[1 + \exp(x(\beta_n - \delta_i))], \tag{9.1}$$

where $x \in \{0, 1\}$ is the dichotomous response variable for person n with ability β_n responding to item i with difficulty δ_i. Two derivations of the Rasch model, the Rating Scale Model (Andrich 1978) and the Partial Credit Model (Masters 1982) have been developed to deal with polytomous items. It can be expressed in a more general form

$$\Pr\{X_{ni} = x\} = \left[\exp(x(\beta_n - \delta_i)) - \sum_{k=1}^{x} \tau_{ki}\right] / \sum_{x=0}^{m_i} \left[\exp(x(\beta_n - \delta_i)) - \sum_{k=1}^{x} \tau_{ki}\right], \tag{9.2}$$

where $x \in \{0, 1, 2 \ \ldots \ m_i\}$ is the integer random variable, and $\tau_{1i}, \tau_{2i}, \tau_{3i}, \ldots,$ are m thresholds between $m + 1$ ordered categories, where m is the maximum score of item i. For any polytomous items, it sometimes occurs that an extreme category, which is logically possible, is not observed in a particular sample (Linacre 2003). For example, in education when the same task performed by the students from different year levels are judged by the same set of criteria, it is likely that none of the lower year level students would achieve the highest mark in some criteria and none of the higher year level students would achieve the lowest mark in some criteria. It could also happen in health. When a group of generally healthy participants are measured by an instrument designed to detect some particular symptoms, it is likely that none of them would achieve the extreme score in some categories.

Collapsing adjacent categories is a common method to treat reversed thresholds and it is theoretically justified when the discrimination of a threshold between two adjacent categories is 0 (Andrich 2009). It also shows that collapsing unobserved categories improves the stability of item estimates (Li 2012). As the threshold shows very poor discrimination between the unobserved extreme category and its adjacent category, collapsing them may be one way to treat the unobserved categories. Although unobserved categories generally can be well dealt with by the principal component reparameterization technique (Andrich and Luo 2003; Luo and Andrich 2005), this paper uses a simulation study to investigate the impact of unobserved extreme categories on item and person estimates under varying number of items with an unobserved extreme category and the feasibility of collapsing unobserved extreme categories as a method to treat unobserved categories in data sets.

9.2 Method

Based on the Partial Credit Model (Masters 1982), data were simulated, using WinGen3 (Han 2006–2012), for 1,000 persons, N ~ (0,1), and ten polytomous items with five categories to investigate two scenarios, one with unobserved extreme high categories and the other with unobserved extreme low categories. The unobserved extreme high category was simulated by changing the value of the last threshold of an item to a large value (personal communication with Mike Linacre), "9" in this case. The unobserved extreme low category was simulated by changing the value of the first threshold of an item to a small value, "-9" in this case. In order to investigate whether different number of items with an unobserved extreme category would affect person and item estimates, three sets of data were simulated under each scenario with 1, 3 and 5 items having an unobserved extreme category. To focus on the matter, in this study, each selected item has only one unobserved extreme category. The generated data sets were analysed with the RUMM2030 software (Andrich et al. 2009). Besides means and standard deviations for person and item measures, mean bias and root mean square error (RMSE) for both person and item measures are reported in this study to compare the estimation accuracy when unobserved extreme categories are present with the accuracy when unobserved extreme categories are collapsed with their adjacent categories. The mean bias measures the difference between the expected value and the true value of a parameter being estimated. It is expressed as $E[\hat{\theta} - \theta]$. When there is less systematic error, the value is close to 0. RMSE incorporates both bias and residual error variance. It is computed as $\sqrt{E[\hat{\theta} - \theta]^2}$. A smaller value of RMSE indicates less estimation error.

9.3 Results

The results for each simulation set under each scenario are presented in details as follows.

Table 9.1 Means and SDs of person and item measures when there is 1 item with an unobserved extreme high category

	1 High Null Cat			
	Original simu	Run 1	No extreme simu	Rescore
Person mean	−0.35	−0.35	−0.11	−0.12
Person SD	1.01	1.08	1.01	1.08
Item mean	0	0	0	0
Item SD	0.56	0.53	0.38	0.38

Table 9.2 Means and SDs of person and item measures when there are 3 items with an unobserved extreme high category

	3 High Null Cat			
	Original simu	Run 1	No extreme simu	Rescore
Person mean	−0.72	−0.60	−0.02	−0.02
Person SD	1.01	1.09	1.01	1.08
Item mean	0	0	0	0
Item SD	0.91	0.76	0.40	0.38

9.3.1 Unobserved Extreme High Categories

Three data sets were simulated with 1, 3 and 5 items having an unobserved extreme high category. The value of the last threshold of an item was changed to a large value of "9", to simulate an unobserved extreme high category. Since there is no change made to person locations, the simulated true person mean and standard deviation are always 0 and 1.01 (accurate to the second decimal place), respectively. Item means and SDs for Original Simu are calculated based on all simulated thresholds including the thresholds with extreme values, "9"s. They are compared with the estimates from Run 1, generated by RUMM2030, which is the first analysis without collapsing the unobserved extreme high categories. In order to be compared with the estimates from Rescore which is the second analysis after collapsing unobserved extreme high categories, item means and SDs for No Extreme Simu are calculated based on all simulated thresholds excluding the thresholds with extreme values, "9"s.

To enable meaningful comparison between the simulated true person and item locations and the estimates generated by RUMM2030, an extra step is required that shifting the simulated true person and item locations to the new origin, set at the simulated item mean. For example, if the simulated item mean is 0.5 and the simulated person mean is 0, after setting the item mean as the origin, the new item mean becomes 0 and person mean becomes −0.5. It is because the mean of the item estimates is fixed as the origin of the scale in RUMM2030.

The means and standard deviations of person and item measures, when there are 1, 3 and 5 items with an unobserved extreme high category, are reported in Tables 9.1, 9.2 and 9.3, respectively. The results showed that all the person SDs for Run 1 and Rescore have similar values (see Tables 9.1, 9.2 and 9.3). When there

Table 9.3 Means and SDs of person and item measures when there are 5 items with an unobserved extreme high category

	5 High Null Cat			
	Original simu	Run 1	No extreme simu	Rescore
Person mean	−1.09	−0.91	0.07	0.08
Person SD	1.01	1.06	1.01	1.06
Item mean	0	0	0	0
Item SD	1.06	0.87	0.38	0.39

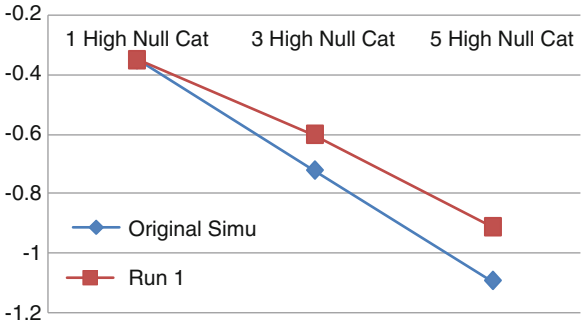


Fig. 9.1 Comparison of person means without collapsing unobserved extreme high categories

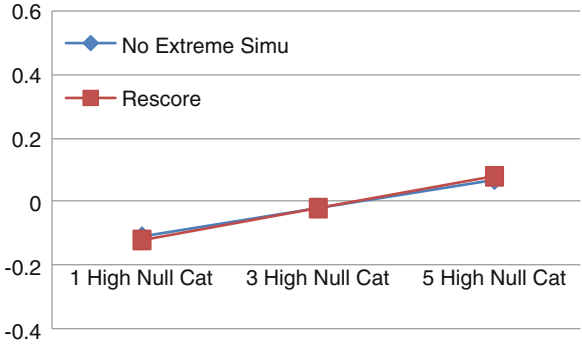


Fig. 9.2 Comparison of person means after collapsing unobserved extreme high categories

was only 1 item with an unobserved extreme high category, item and person means and SDs from Run 1 were very close to the simulated values (see Table 9.1). The estimates from Rescore after collapsing the unobserved extreme high category also performed well. As the number of items with an unobserved extreme high category increased, the person mean and item SD from Run 1 showed somewhat discrepancy when compared with the simulated values (see Tables 9.2 and 9.3). The person means for Run 1 were greater than the simulated values and the item SDs are less than the simulated values (see Figs. 9.1 and 9.3). However, the estimates from Rescore were still very close to the simulated values (see Figs. 9.2 and 9.4).

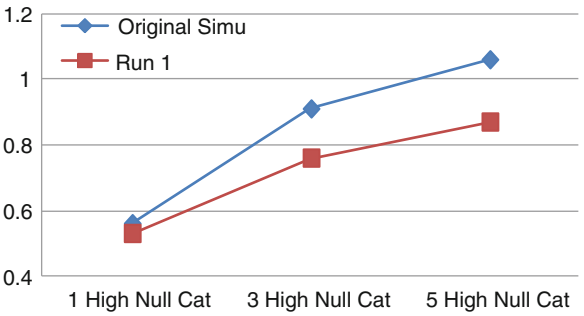


Fig. 9.3 Comparison of item SDs without collapsing unobserved extreme high categories

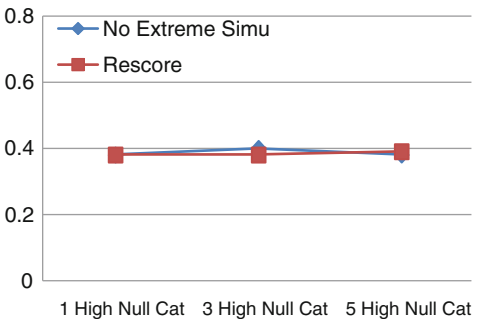


Fig. 9.4 Comparison of item SDs after collapsing unobserved extreme high categories

Table 9.4 Means and SDs of person and item measures when there is 1 item with an unobserved extreme low category

	1 Low Null Cat			
	Original simu	Run 1	No extreme simu	Rescore
Person mean	0.02	0.00	−0.22	−0.22
Person SD	1.01	1.10	1.01	1.10
Item mean	0	0	0	0
Item SD	0.72	0.66	0.28	0.27

9.3.2 Unobserved Extreme Low Categories

Three simulation sets were generated to simulate 1, 3 and 5 items with an unobserved extreme low category. Similarly, the value of the first threshold of an item was changed to a small value of "-9" to simulate an unobserved extreme low category. The means and standard deviations of person and item measures, when there are 1, 3 and 5 items with an unobserved extreme low category, are reported in Tables 9.4, 9.5 and 9.6, respectively. Similar to the results from unobserved

Table 9.5 Means and SDs of person and item measures when there is 3 items with an unobserved extreme low category

	3 Low Null Cat			
	Original simu	Run 1	No extreme simu	Rescore
Person mean	0.40	0.32	−0.33	−0.32
Person SD	1.01	1.08	1.01	1.07
Item mean	0	0	0	0
Item SD	0.96	0.81	0.39	0.39

Table 9.6 Means and SDs of person and item measures when there are 5 items with an unobserved extreme low category

	5 Low Null Cat			
	Original simu	Run 1	No extreme simu	Rescore
Person mean	0.81	0.67	−0.40	−0.41
Person SD	1.01	1.07	1.01	1.07
Item mean	0	0	0	0
Item SD	1.01	0.85	0.43	0.46

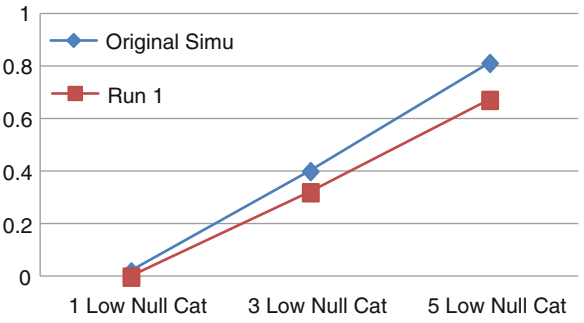


Fig. 9.5 Comparison of person means without collapsing unobserved extreme low categories

extreme high categories, person SDs for Run 1 and Rescore were almost identical (see Tables 9.4, 9.5 and 9.6). Collapsing the extreme low categories also resulted in more accurate person means and item SDs than the original analysis when the proportion of the items with an unobserved extreme low category increased (see Figs. 9.5, 9.6, 9.7 and 9.8).

9.3.3 Bias and RMSE

In order to investigate the deviations of person and item estimates, the bias and RMSE statistics are also reported in this study (see Tables 9.7 and 9.8). For all six data sets, the value of item bias is 0. It is because, as mentioned earlier, the mean of item measures is arbitrarily set to be "0" to enable meaningful comparisons.

Fig. 9.6 Comparison of person means after collapsing unobserved extreme low categories

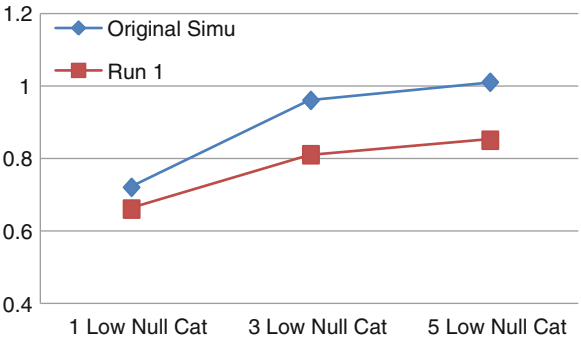


Fig. 9.7 Comparison of item SDs without collapsing unobserved extreme low categories

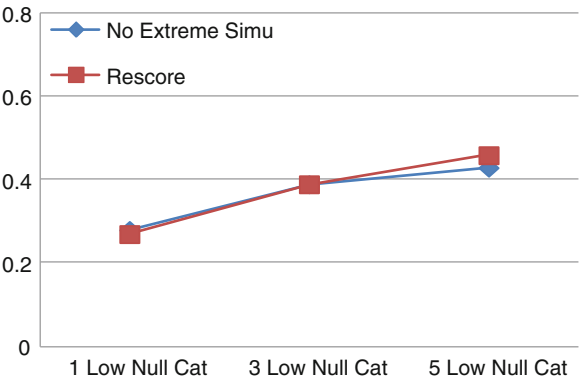


Fig. 9.8 Comparison of item SDs after collapsing unobserved extreme low categories

Table 9.7 showed that collapsing unobserved extreme categories generated satisfactorily unbiased person measures. When the number of items with an unobserved extreme category increased, the person estimates from the first analysis became more biased. Consistent with the results from the comparison of person means, unobserved extreme high categories seemed to systematically overestimate person

Table 9.7 Bias of person and item measures for non-collapsed and collapsed unobserved extreme categories

	Person bias		Item bias	
Data set	Run 1	Rescore	Run 1	Rescore
1 High Null Cat	0.00	−0.01	0.00	0.00
1 Low Null Cat	−0.02	0.01	0.00	0.00
3 High Null Cat	0.12	0.00	0.00	0.00
3 Low Null Cat	−0.07	0.01	0.00	0.00
5 High Null Cat	0.18	0.01	0.00	0.00
5 Low Null Cat	−0.14	0.00	0.00	0.00

Table 9.8 RMSE of person and item measures for non-collapsed and collapsed unobserved extreme categories

	Person RMSE		Item RMSE	
Data set	Run 1	Rescore	Run 1	Rescore
1 High Null Cat	0.39	0.39	0.02	0.02
1 Low Null Cat	0.38	0.38	0.03	0.01
3 High Null Cat	0.42	0.39	0.23	0.01
3 Low Null Cat	0.39	0.38	0.15	0.01
5 High Null Cat	0.43	0.39	0.23	0.01
5 Low Null Cat	0.42	0.39	0.17	0.01

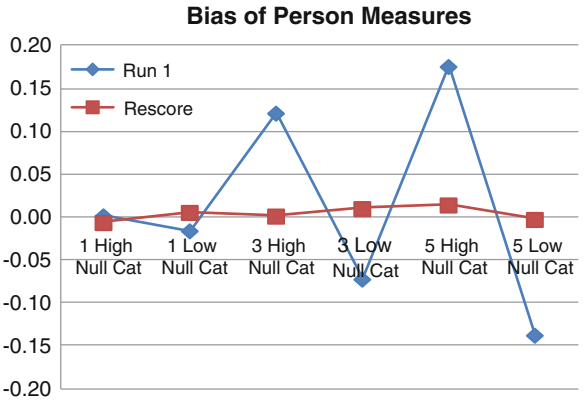


Fig. 9.9 Bias of person measures for non-collapsed and collapsed unobserved extreme categories

measures, a positive bias, and unobserved extreme low categories tended to systematically underestimate person measures, a negative bias (see Fig. 9.9).

Table 9.8 summarized RMSE for person and item measures from 6 data sets and showed that the impact of the unobserved extreme categories on the person estimates was smaller than the impact on the item estimates. There seemed to be more impact when the proportion of the items with an unobserved extreme category increased. Collapsing unobserved extreme categories resulted in significantly smaller RMSEs of item measures (see Figs. 9.10 and 9.11).

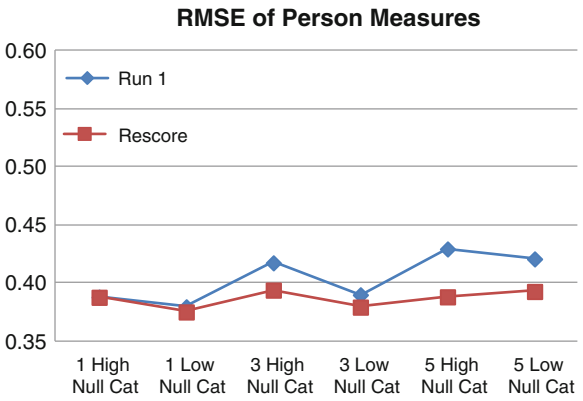


Fig. 9.10 RMSE of person measures for non-collapsed and collapsed unobserved extreme categories

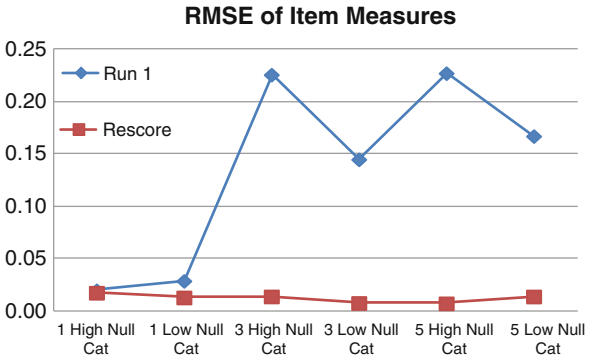


Fig. 9.11 RMSE of item measures for non-collapsed and collapsed unobserved extreme categories

9.4 Conclusion and Discussion

This study investigated the impact of unobserved extreme categories on the person and item measures and proposed a possible solution by collapsing unobserved extreme categories. Two scenarios, the presence of unobserved extreme high categories and unobserved extreme low categories, were examined. It was discovered that unobserved extreme categories did have an impact on person and item estimates and the impact became stronger as the number of items with an unobserved category increased.

In the first scenario, the presence of unobserved extreme high categories resulted in an overestimated person mean. The mean bias statistics for the person measures suggested that the presence of unobserved extreme high categories tended to generate the person estimates that are systematically higher than the simulated person locations. In the second scenario, unobserved extreme low categories led to an underestimated person mean. The mean bias statistics for person measures showed that the presence of unobserved extreme low categories had a tendency to produce the person estimates that are systematically lower than the simulated

person locations. As a solution, collapsing unobserved extreme categories successfully corrected the systematic error in person estimates and produced satisfactory person mean estimates. There was little sign of systematic error in item estimates as mean bias was 0 (accurate to the second decimal place) for both the estimates with unobserved extreme categories and the estimates with unobserved extreme categories collapsed.

In both scenarios, collapsing unobserved extreme categories did not make much difference in person SDs. The RMSEs of the person estimates with unobserved extreme categories were slightly greater than those with unobserved extreme categories collapsed. This seems to be caused by the bias that existed in the person estimates with unobserved extreme categories. However, unobserved extreme categories seemed to have a great impact on the item estimates, which resulted in smaller item SDs than the simulated values. The RMSE statistics also suggested that there were some significant deviations from the simulated item locations especially when a number of items had an unobserved extreme category. There was more severe measurement error when estimating the thresholds associated with an unobserved extreme category. Collapsing unobserved extreme categories satisfactorily recovered the simulated item locations, as the RMSEs were very close to 0.

In summary, the results suggest that collapsing unobserved extreme categories improves both person and item estimation accuracy, especially when there are a relatively large number of items with an unobserved extreme category. In general, unobserved extreme high categories were found to lead to greater mean bias in person estimates and greater RMSE in item estimates than do unobserved extreme low categories. These findings have implications for designing and measuring performance tasks that need to be carried out across a wide spectrum of ability groups, in which case mis-targeting is inevitable. For example, pooling up all the data from various ability groups and calibrating them as a whole, which masks the existence of unobserved extreme categories for some ability groups, may result in less accurate person measures than calibrating the data separately based on ability groups. It may also affect the common-item equating procedure for polytomous items where the threshold values of common items are used. Severe deviations from true threshold locations caused by unobserved extreme categories would lead to obscured and confounding equating results.

References

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

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

Andrich, D. (1989). Distinction between assumptions and requirements in measurement in the social sciences. In *Proceedings of the XXIVth international congress of psychology* (Vol. 4, Mathematical and theoretical systems, pp. 7–16). B.V. North Holland: Elsevier Science Publications.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1 suppl), I-7–I-16.

Andrich, D. (2009). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications*. Hillsdale: Lawrence Erlbaum Associates, Inc. Chapter 6.

Andrich, D., & Luo, G. (2003). Conditional estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement, 4*, 205–221.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2009). *RUMM 2030 [Computer Software]*. Perth: RUMM Laboratory.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Lawrence Erlbaum Associates.

Duncan, O. D. (1984). Rasch measurement: Further examples and discussion. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 367–401). New York: Russell Sage.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457–459.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessment, 13*(12).

Kuhn, T.S (1961/1977). The function of measurement in modern physical science. *Isis, 52,* 161–190. Reproduced in T.S. Kuhn (1977), *The Essential Tension*. Chicago: The University of Chicago Press.

Li, E. F. (2012). Convergence, collapsed categories and construct validity. *Rasch Measurement Transaction, 25*(4), 1339–1340.

Linacre, J. M. (2003). Unobserved categories: Estimating and anchoring Rasch measures. *Rasch Measurement Transaction, 17*(2), 924–925.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75*(2), 127–132.

Luo, G., & Andrich, D. (2005). Estimating parameters in the Rasch model in the presence of null categories. *Journal of Applied Measurement, 6*(2), 128–146.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Expanded edition 1980. Chicago: University of Chicago.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333). Berkeley: University of California Press.

Tennant, A., & Conaghan, P. (2007). The Rasch measurement model in rheumatology: What is It and why use it? When should It be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism, 57*(8), 1358–1362.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*(Suppl 1), S22–S26.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*, 529–554.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857–860.

Wright, B. D., & Stone, M. (1999). *Measurement essentials*. Wilmington: Wide Range INC.

Chapter 10
Assessment Report on Reading Literacy in Guangxi Ethnic Minority Region—Based on PIRLS 2006 Test Analysis

Jing Yu and Dehong Luo

Abstract On the basis of the statistic analysis of the result of PRILS2006 assessment of 860 Grade-3-or-4 pupils from 37 classes, 12 elementary schools, 9 cities in Guangxi Ethnic Minority Region, the following conclusions have been made: Guangxi average score is lower than the international average; the score distribution is "bi-little and bi-large"; female students are better than males; Han nationality students perform the poorest; In-school Chinese test score (X_1) and family BC (X_2) are two predictable factors of the reading literacy performance(Y) ($Y = 373.790 + 30.821X_1 + 11.754X_2$), but it can't account for the performance of both the high score group and the low score group; Mothers tend to break the set pattern of the occupational hierarchy; family material condition is paid more attention to than books; about 30-min free reading per day and collection of 26–200 books can mostly help pupils reach the intermediate level of reading literacy; reading capacity cultivated in Chinese classes is different from PIRLS reading literacy. The following suggestions are put forward: develop pilot project of continuous assessment and monitor of reading literacy; carry out the local experiment of PIRLS reading literacy to provide scientific data for international comparative analysis; make use of library resources to improve the balanced degree of reading chance as an aid to enhance balanced development of education.

Keywords Guangxi PRILS2006 test reading literacy experiment • Assessment and monitor balanced reading chance

J. Yu • D. Luo (✉)
Educational Department, College of Education, Guangxi University, NO.100 East University Road, Guangxi Zhuang Autonomous Region, Nanning 530004, People's Republic of China
e-mail: 517792580@qq.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_10,

10.1 Assessment Background

During the past years, there has been a great tendency all over the world to concern educational quality and assess scientifically students' academic performance, among which reading literacy assessment is the indispensible one and has been comparatively studied internationally. Two of the most influential ones are PIRLS (Progress in International Reading Literacy Study) and PISA (Program of International Student Assessment) separately organized by IEA (The International Association for the Evaluation of Educational Achievement) and OECD (Organization of Economic and Cooperation and Development). Both of them evaluate from the view of cognition how well the students master the knowledge and skill appreciated and required by themselves, and investigate in what extend the personal, family, school and government factors influence the reading literacy performance. In 2009, Shanghai, as one of the most developed cities in China, participated in PISA2009 Reading, Science and Mathematics Assessment, and got the first place in Reading Literacy. Then, what about the Guangxi Zhuang Autonomous Region, inhabited by various minority groups, and one of the least undeveloped areas in China? That's the aim of this research to find out the gap of educational quality and enhance the development of China's educational equity policy.

10.2 Assessment Content and Methods

In April, 2012, Jieli Level Reading Research Centre, under management of Jieli Publishing House, organized 860 pupils to participate in PIRLS2006 Assessment. The pupils are from 37 classes of 12 primary schools in Guangxi 9 cities, including Nanning, Liuzhou, Guilin, Beihai, Baise, Hechi, Qinzhou, Wuzhou and Hezhou. Until now, Guangxi is the first place in China to take part in the test of PIRLS.

10.2.1 Assessment Tools

The reading literacy test items and scoring criterion are the Chinese Version, adopted by Hong Kong and Taiwan. The original questionnaires are somewhat modified and selected according to the specific circumstances, with pupils as the only respondents, including 6 aspects and 16 items, such as basic information (age, gender and nationality), parents' occupational hierarchy (OH), family conditions, personal reading time, family collection of books and In-school Chinese exam score (SCES), required finishing in 10 min.

10.2.2 Sampling Methods

Sampling method is internationally applied binary stratified random sampling, ensuring certain representativeness. At the first stage, Guangxi primary schools are classified into three kinds, i.e. provincial capital schools, other city schools and town schools, and each kind of sampling school covers different educational quality schools, from which 12 schools are randomly sampled, pro rata with the number of the kind of the school. At the second stage, 860 pupils from 37 classes are randomly sampled from the 12 schools.

10.2.3 Raw Data Processing

Since Guangxi isn't the member of IEA, nor it has applied for the participation of PIRLS test, raw data hasn't been sent to IEA Data Processing Centre for treatment. In order to make our data comparable with that of the other 45 countries and areas, we have read up on PIRLS2006 Technical Report before we use the similar method to IEA to process the raw data and transform standard scores with the tools of SPSS17.0 and EXCEL according to the basic and conventional formulas of educational statistics and measurement.

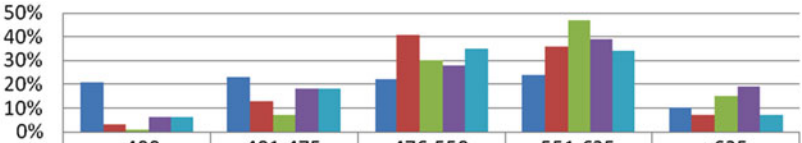
10.3 Transnational and Cross-Border Analysis of Guangxi Reading Literacy

10.3.1 The Overall Level Is Low

Among the 45 participating countries and areas, Guangxi average score is 496 (SE5.0), in 36th place, lower than the international average score 500 and 69 points lower than the first place country Russia; Among the 8 Asian participators, Guangxi score is in the 4th place, lower than Hong Kong (564), Singapore (558) and Taiwan (535), but higher than Iran (421), Indonesia (405), Kadar (353) and Kuwait (330) (PIRLS 2006).

10.3.2 The Score Distribution Is "Bi-little and Bi-large"

PIRLS uses four points on the scale as international benchmarks. The benchmarks represent the level of performance shown by students internationally. For PIRLS 2006, the Advanced International Benchmark is over 625, the High International Benchmark is 550–624, the Intermediate International Benchmark is 475–549, and the Low International Benchmark is 400–474. According to it, the score distribution of Guangxi pupils is characterized by "Bi-little and Bi-large".



	<400	401-475	476-550	551-625	>625
Guangxi	21%	23%	22%	24%	10%
Taiwan	3%	13%	41%	36%	7%
HongKong	1%	7%	30%	47%	15%
Singapore	6%	18%	28%	39%	19%
International	6%	18%	35%	34%	7%

Fig. 10.1 Cross-border analysis of international benchmark

"Bi-little" means the little number of the students who reach the advanced benchmark (10 %) and high benchmark (34 %), compared with Hong Kong (15 %) and Singapore (19 %), although larger than Taiwan (7 %) and the international average (7 %) (Fig. 10.1). "Bi-large" means the large number of the students (21 %) below low benchmark (<400) and the large standard error (5.0). As far as the former is concerned, the relative number is 3 % in Taiwan, 1 % in Hong Kong, 6 % in Singapore and the 6 % internationally. In terms of standard error, Guangxi is in the third place internationally, as large as 5.0, meaning the obvious discrete distribution score, which implies that the differences in reading literacy achievement among students are very great.

10.3.3 The Level of Comprehension Achievement in Interpreting Is Lower Than That in Straightforward

According to PIRLS2006, reading is a cognitive process, including four reading literacy, such as retrieving, straightforward inference, interpreting and integrating, and evaluating. The former two, called as straightforward comprehension, is an automatic thinking, while the latter two, called as interpreting comprehension, requires readers to be based on their own knowledge to construct a deep comprehension and critical thinking of the text. Guangxi participants get higher score in straightforward comprehension (507) than in interpreting understanding (498), opposite to participants from Hong Kong (558 Vs. 566) and America (532 Vs. 546). As far as the items for evaluating literacy, Guangxi pupils have got 0.47 average points out of 4 (raw score), while in terms of items for interpreting comprehension, the pass rate is only 0.32, much lower than Taiwan (0.49), Hong Kong (0.60) and Singapore (0.57), which means that Guangxi pupils find the evaluating items more difficult than the other Asian ones. It can be concluded that Guangxi pupils are at a low level thinking ability with poor analysis, inference and critical thinking.

Table 10.1 Gender difference in reading literacy in Guangxi

	Total score	Retrieving	Straight forward	Interpreting and integrating	Evaluating	Literary text	Informational text
Boy	492.6	493.5	493.2	497.8	489.3	498.8	489.4
Girl	512.3	512.2	510.1	507.5	513.4	507.8	513.5
D	19.7	18.7	16.9	9.7	24.1	8.7	24.1

Table 10.2 Nationality difference in reading literacy

	Total score	Retrieving	Straight forward	Interpreting and integrating	Evaluating	Literary text	Informational text
Han	497.3	500.2	492.4	506.0	495.4	497.5	497.7
Zhuang	514.2	508.0	517.1	504.4	518.1	514.2	511.9
Other	518.9	508.5	547.4	494.3	482.0	520.5	514.7
D	21.6	8.3	55	11.6	13.4	23	17

10.3.4 Girls Do Better Than Boys and Han Nationality Pupils Perform the Poorest on the Whole

The gender ratio of Guangxi boys and girls is 45:54, and girls do better in all the reading literacy than boys (Table 10.1), in accord with the international situation. T-test shows that, apart from evaluating literacy ($t = -2.373$, $p = 0.018 < 0.05$) and informational text ($t = -2.357$, $p = 0.019 < 0.05$), other indexes lack the significant level, while internationally, the difference between girls (509) and boys (492) has reached the significant level (Ke Huawei et al. 2009).

In terms of nationality, participants are divided into Han, Zhuang and other ethnic minority (Yao, Dong, Miao, Maonan, etc.). Han Nationality pupils get the lowest score in all the achievement, while other ethnic minority pupils get the highest (Table 10.2). There isn't correlation between nationality and reading literacy, gender, age, family condition and mother's OH, whereas a weak correlation ($r = 0.178$) between nationality and father's OH. It can be concluded that nationality isn't the impact factor of reading achievement; hence the national equality policy has been executed well in Guangxi.

10.4 Analysis of Factors That Influence Reading Literacy

Factors that are supposed to influence reading literacy are classified into three kinds, which are family background, including parents' OH, family condition and family collection of books; personal free reading time per day; self-report in-school Chinese Language exam score.

Table 10.3 Percent distribution of parents' OH

	Others	Private enterprises	Executives	Professionals	National and community managers
Father	58.5	6.4	8.6	11.1	6.4
Mother	68.1	4.7	2.0	12.6	2.7

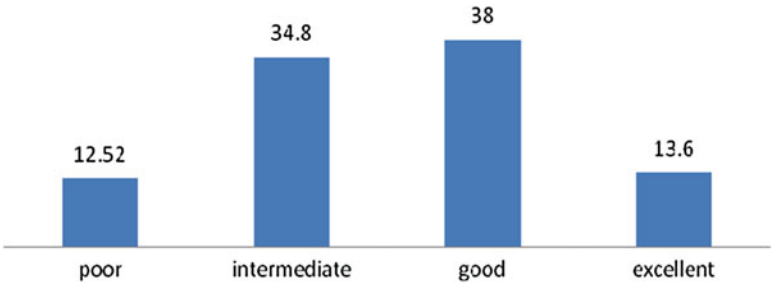


Fig. 10.2 Percentage distribution of family condition

10.4.1 Analysis of Family Background Factor

In the questionnaires parent's occupation is classified into five hierarchies, that is, the lowest hierarchy, others (clerk, individual business, commercial service, industry, agriculture, urban and rural unemployed and half-unemployed), private enterprises, executives, professionals, and national and community managers (Lu Xueyi 2002). Investigation shows that the ratio of each hierarch (Table 10.3) is consistent with the ratio and its developing trend since China's reform and opening-up (Lu Xueyi 2010). In other words, sampling schools and pupils are representative.

Family condition covers learning facilities and environment, including desk, private room, quiet learning place, learning computer, learning software, internet, calculator, classic literature books. The result is leveled as four classes, i.e. poor, intermediate, good and excellence, showing a bell curve distribution diagram (Fig. 10.2), telling us that, generally speaking, the pupils have good learning environment.

The result of family BC is divided into six levels and shows a skewed normal distribution, with 33.6 % pupils having less than 25 books. It means that books are paid less attention to than other learning facilities (Fig. 10.3).

In statistic analysis (Table 10.4), in a descending order, the correlation coefficient between reading literacy and family background is family condition ($r = 0.3$, $p < 0.05$), family BC ($r = 0.22$, $p < 0.05$) and father OH ($r = 0.15$, $p < 0.05$); Significant difference test level is family BC ($T = 57.247$), family condition ($T = 43.251$), father's OH ($T = 26.904$) and mother's OH ($T = 25.690$); Eta squared is family condition (0.092), family BC (0.061), father's OH (0.032) and

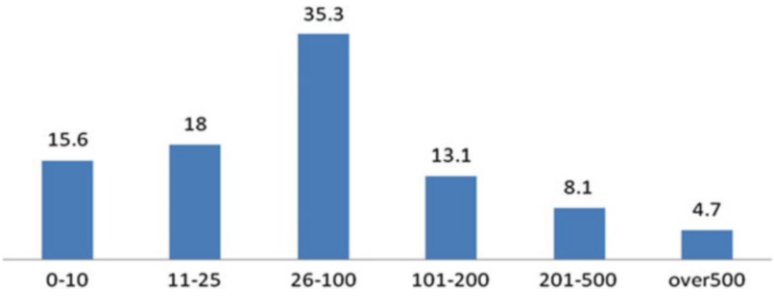


Fig. 10.3 Percentage distribution of family book collection

Table 10.4 Family background as factors to influence reading performance (95 % confidence interval)

	Parameter estimation R	Std E	T value	P value	Eta squared
Father's OH	0.15	.071	26.904	0.02	0.032
Mother's OH	0.089	.064	25.690	0.4	0.011
Family condition	0.3	.044	43.251	0.02	0.092
Family BC	0.22	.068	57.247	0.00	0.061

mother's OH (0.011). That is to say, family BC is the most significant factor to influence reading literacy achievement, and family condition change will mostly result in reading achievement with other situations unchanged. Each index related to mother lies the last one, which in fact tells that mother is more inclined to smash the stereotype that OH decides children's learning performance.

Participants with what kinds of family background have the best reading achievement? Our result is the following. The better the family condition is, the higher score the child gets, with differences as high as over 100 points. The participants whose family BC is over 500 have the best reading achievement, and the next two are those with 26–100 books and 100–200 books, which is consistent with other researches (NIES 2011); The participants whose parents are professionals get the highest score, and the successive descending order is national and social manager, executive, others and private enterprise (Table 10.5).

10.4.2 Analysis of Free-Reading-Time-Per-Day Factor

Free time reading is a kind of non-textbook autonomous reading. Results show that 3.2, 56.3, 14.3 and 20.7 % pupils separately choose to read minutes of zero, 0–30, 30–60 and over 60, and the corresponding scores they get are 467, 500.3, 497.4 and

Table 10.5 Literacy score based on family background factors

Father OH	Literacy	Mother OH	Literacy	Family BC	Literacy	Family condition	Literacy
Others	497.8	Others	504.2	0–10	459.7	Poor	440.1
Business enterprise	508.4	Business enterprise	486.5	11–25	488.5	Intermediate	488.0
Executive	501.2	Executive	528.8	26–100	518.0	Good	516.8
Professional	549.7	Professional	525.6	101–200	517.8	Excellent	546.1
National manager	527.8	National manager	536.2	201–500	511.9		
				Over 500	556.8		

Table 10.6 Influence of reading time and SCTS (95 % confidence interval)

	Parameter estimation R	Std E	T value	P value	Eta squared
Reading time per day	0.112	.044	57.559	0.00	0.017
SCTS	0.278	.0448	71.116	0.00	0.102

525.9. Specifically, pupils with over 60 min' reading are far ahead, especially in the literary text and evaluating literacy. Deserve to be mentioned, the second best achievement belongs to the pupils with less than 30 min' reading. However, although they do well in the literacy of straightforward inference and retrieving, they are poor in interpreting and integrating literacy.

Analysis shows that the correlation between reading literacy and reading time is weak significant ($r = 0.11$, $p < 0.05$), but its standard error is rather low and T value is higher than family background (Tables 10.4 and 10.6). It means that differences in pupils' reading interest and devotion influence reading level more than family background factors.

10.4.3 Analysis of Factor of Self-Reported SCTS

The factor of SCTS aims to discover the relation between existing Chinese classroom teaching and reading literacy. The reporting result marks the percentage distribution of poor (lower than 60), intermediate (60–79), good (80–89) and excellent (over 90) performance is successively 3.2, 6.4, 19.8 and 43.7 %, constituting a set of going-up stairs instead of normal distribution, quite inconsistent with the distribution of PIRLS2006 reading literacy (Fig. 10.1).

Statistic analysis shows that the correlation between SCTS and reading literacy is significant ($r = 0.278$, $p < 0.05$), and standard error is rather low, but T value is the highest, as high as 71.116. It can be concluded that SCTS is the most influencing factor.

Table 10.7 Reading literacy based on SCTS

SCTS	Total score	Retrieving	Straight forward	Interpreting and integrating	Evaluating	Literary text	Informational text
Poor	485.9	499.5	509.8	453.2	468.1	481.5	491.1
Intermediate	453.9	454.3	452.2	488.3	450.5	457.9	457.9
Good	481.3	484.0	480.0	488.9	483.4	483.4	483.2
Excellent	540.6	535.0	535.5	535.5	533.7	536.3	537.4

Table 10.8 Model summary

Model	R	R square	Adjusted R square	Std. error of the estimates
1.	.283[a]	.080	.077	98.821
2.	.320[b]	.102	.096	97.811

Dependent variable: total score
[a]Predicator: (constant), SCTS
[b]Predicator: (constant), SCTS, book collection

Table 10.9 Coefficient

	Unstandardized coefficients		Standardized coefficient	t		
Model	B	Std. error	Beta	Sig.		
1.	(constant)	395.278	25.238		15.662	.000
	SCTS	34.940	7.119	.283	4.908	.000
2.	(constant)	373.790	26.318		14.203	.000
	SCTS	30.821	7.223	.250	4.267	.000
	Book collection	11.754	4.531	.152	2.594	.010

Dependent variable: total score

Although of course the pupils with excellent SCTS do the best in reading literacy, quite meaningfully, the second best literacy performance, with 32 points higher than the third one, belongs to those with below-60 SCTS. Their reading dominance is straightforward comprehension and informational text, while interpreting comprehension is rather poor (Table 10.7). It means that the reading ability that Guangxi existing Chinese classroom teaching aims to cultivate is different from what PIRLS2006 proposes.

10.4.4 Multiple Linear Regression Analysis

All the factors are tested with the multiple linear regression model, with the total score as the dependent variable, and gender, nationality, OH, family condition, family BC, reading time per day and SCTS as the independent variables, resulting in two model summary (Table 10.8) and coefficients (Table 10.9). Results reveal that SCTS and family BC are two mainly predicating factors.

10.5 Analysis of High Score Group (HSG) and Low Score Group (LSG)

Participants are classified into HSG and LSG according to percentiles. Those above 75 percentile are HSG and those below 25 percentile are LSG, with respectively 583.26 and 419.90 as the break point. The reason that we don't adopt the international benchmark is that the gap between 10 % high scorer (over 625) and the 21 % low scorers (below 400) is too wide.

In terms of gender, 32 % male high scorers are only as half as 68 % female high scorers, but boys are better than girls in each kind of performance, although with no significant difference except in literary text; for LSG, the ratio between boys and girls is 46 % and 54 %, with no significant difference and large score gap (Table 10.10). It can be concluded that although there are fewer male high scorers than female, they are more excellent than girls, and girls' scores are more evenly distributed. That's to say, boys are both the best and the poorest.

In terms of parents' OH, considering the percentage distribution is huge different, especially with 58.5 % father and 68.1 % mother consisting of the first OH, we think it more strategy to analyze the percentage of different OH in high and LSG (Table 10.11). In HSG, professionals consist of the most and the next is national managers, and least contribution is business enterprises and others. In LSG, others and executives consist of the most, and the least contribution is professionals and national managers. Statistics tell us that the children of professionals and national

Table 10.10 Gender analysis of high and LSG

	Total score	Retrieving	Straight forward	Interpreting and integrating	Evaluating	Literary text	Informational text	
HSG	Boy	636.9	604.7	617.0	637.3	618.5	622.8	626.0
	Girl	624.5	591.57	615.1	611.5	610.8	596.3	626.0
LSG	Boy	374.5	385.47	374.7	411.4	414.0	388.6	382.3
	Girl	380.0	385.47	377.5	422.8	425.8	394.7	387.4

Table 10.11 Analysis of factors of family background in high and LSG

OH	Others	Business enterprises	Executives	Professionals	Managers	
HSG (%)	Father	23	31	26	44	35
	Mother	26	16	50	33	45
LSG (%)	Father	27	19	26	11	19
	Mother	26	26	25	22	18

Family condition	Poor	Intermediate	Good	Excellent
HSG (%)	2.9	24.5	48.0	24.5
LSG (%)	23.9	38.9	30.1	7.1

Family BC	0–10	11–25	26–100	101–200	201–500	>500
HSG (%)	5.9	16.8	42.6	15.8	9.9	8.9
LSG (%)	25.5	25.5	25.5	13.7	8.8	1.0

Table 10.12 Analysis of reading time and SCTSs in HSG and LSG

Reading time (min)		0	0–30	30–60	>60
HSG (%)		1.0	56.4	11.9	30.7
LSG (%)		4.0	62.4	14.9	18.8
SCTSs	<60	61–79		80–89	>90
HSG (%)	2.2	4.4		15.6	77.8
LSG (%)	5.6	15.3		37.5	41.7

managers are the most advantageous of being excellent, but they don't have any advantage of free from low score.

In terms of family condition, excellent condition consists of 24.5 % in HSG and 7.1 % in LSG, while good condition consists of 48 % in HSG and 30.1 % in LSG. It can be said that family condition is favorable for pupils to reach excellent performance (Table 10.11).

From the point of family BC, families with fewer than ten books contribute to 6 and 26 % respectively in high and LSG. In LSG, families with books below 25 consist of as high as 50 %, while in HSG, nearly 60 % families have 26–100 books. Statistics shows that it's unfavorable for participants with books below 25 to reach basic benchmark, and favorable for participants with books over 500 to reach advanced and excellent benchmark (Table 10.11).

As far as reading time is concerned, 99 % high scorers and 94 % low scorers read per day. However, one third of high scorers read over 60 min per day. So, reading time isn't an influencing factor for both high and low scorers of literacy performance, but no doubt that the longer he reads, the more probable that he gets high scores. It's very important for pupils to read about 30 min per day (Table 10.12).

For SCTSs, participants with over 90 points consist of 77.8 % high scorers, among which 95 % scorers are over 625, and 41.7 % are low scorers (Table 10.12). We can draw a conclusion that self-reported SCTSs aren't identical with PIRLS2006 reading literacy performance. The vast majority of high scorers have high SCTSs, while low scorers can also reach the average of 85 SCTS.

Statistics analysis (Table 10.13) shows that, neither high scores nor low scores have significant correlation with all the factors. Significance test of difference indicates, in HSG, that the top three differences are SCTSs ($T = 52.672$), family condition ($T = 38.021$) and gender ($T = 35.834$), with SCTS as the most influencing factor; while in LSG, they are gender ($T = 31.3$), SCTSs ($T = 30.3$), reading time per day ($T = 29.6$), with gender as the greatest influencing factor. For Eta square value, in HSG, the top three are SCTSs (0.05), family BC (0.046) and reading time per day (0.037); while, in LSG, they are father's OH (0.05), reading time per day (0.04) and SCTSs (0.04). It can be concluded, in HSG, that the change in SCTS will have a greatest impact on the change in reading literacy with no other factors unchanged; while in LSG, that's the factor of father's OH. We'd like to repeat that mother is more inclined to smash the stereotype that OH decides children's learning performance.

Regression analysis announces that, in HSG (Table 10.14), gender is the only variable to predicate reading literacy, while in LSG (Table 10.15), father's OH is the only predicating variable.

Table 10.13 Analysis of factors influencing high and LSG (95 % confidence interval)

	Parameter estimation R		Std error		T value		P value		Eta squared					
	High		Low		High		Low		High		Low		High	Low
Gender	−0.12		/		0.05		0.05		35.80		31.30		0.13	0.35	0.02	0.01
Nationality	−0.18		/		0.05		0.05		27.90		28.20		0.21	0.78	0.78	0.00
Father's OH	−0.05		0.20		0.15		0.12		14.90		13.10		0.80	0.37	0.02	0.05
Mother's OH	0.01		−0.08		0.14		0.12		13.40		13.10		0.82	0.61	0.02	0.03
Family condition	0.09		−0.07		0.04		0.08		38.00		26.40		0.84	0.66	0.01	0.02
Family BC	0.10		0.08		0.13		0.13		29.60		20.00		0.48	0.87	0.05	0.02
Reading time	−0.09		0.08		0.09		0.08		29.80		29.60		0.30	0.31	0.04	0.04
SCTS	0.18		0.01		0.07		0.10		52.70		30.30		0.22	0.42	0.05	0.04

Table 10.14 Model summary and coefficients for HSG

Model summary[a]

					Change statistics				
Model	R	R square	Adjusted R square	Standard error of estimation	R square change	F change	df1	df2	Sig. F change
1	.239[b]	.057	.046	36.833	.057	5.022	1	168	.028

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. error	Beta	t	Sig.
1	(Constant)	667.643	15.587		42.834	.000
	性别	−19.577	8.843	−.237	−2.214	.030

[a] Dependent variable: total score
[b] Predicator: (constant), gender

Table 10.15 Model summary and coefficients for LSG

Model summary[a]

					Change statistics				
Model	R	R square	Adjusted R square	Std. error of the estimate	R square change	F change	df1	df2	Sig. F change
1	.285[b]	.081	.065	31.474	.081	5.028	1	114	.029

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. error	Beta	t	Sig.
1	(Constant)	361.469	7.005		51.604	.000
	Father's occupational hierarchy	7.024	3.133	.285	2.242	.029

[a] Dependent variable: total score
[b] Predictors: (constant), father's occupational hierarchy

10.6 Assessment Results

10.6.1 Participants' Reading Literacy

The overall reading literacy achievement is lower than international average, leveled as the bottom 11th internationally, quite inconsistent with the self-reported SCTS.

One fifth participant hasn't reached the basic international benchmark (<400) and variance between pupils is quite large.

Interpreting comprehension is lower than straightforward comprehension, with wide poor evaluating literacy.

Girls do better than boys as a whole, but male high scorers are more than female.

As a whole, Han nationality participants do the poorest, and other ethnic minority (Yao, Dong, Miao, Maonan, etc.) participants do the best.

10.6.2 Influencing Factors

Regression analysis shows different predicators of reading performance. For overall reading performance, they are SCTS and family BC, with 95 % predictability; for high and LSG, they are gender and father's OH respectively, with 57 and 81 % predictability.

Mothers are more inclined to smash the stereotype that OH decides children's learning performance.

Participants' family condition is generally good, but books aren't paid enough attention to.

About half an hour's free reading time per day and 26–100 books are very favorable for pupils to reach intermediate international benchmark.

Participants with reading time below 30 min are poor in interpreting comprehension and literary text.

10.7 Suggestion

As discussed above, the two predicators are SCTS and family BC. Since we regard the former as educational quality, and the latter family background factor, hence based on it, the following suggestions are put forward.

10.7.1 Launching Pilot Assessment and Monitor of Reading Literacy to Improve Balanced Development of Education

American psychologist Thorndike thought reading has always been the most widely studied subject (Wright and Stenner 1998) and also the means to be trained and ends to be realized in education. Since reading isn't a hereditary quality, it can't be transformed to learning ends and leaning means unless through vocabulary and syntax learning. In America, reading tops the basic educational level, and reading ability is thought as national power. In China, national reading level is thought to decide national spirit state (Zhu Yongxin 2012). Since reform and opening-up in China, great achievement in Chinese teaching reform has been reached, but the systematic monitor and assessment of Chinese students' reading literacy has always been ignored.

Since the uniform exam of entrance of middle school was abandoned, the unified assessment and monitor of Chinese educational quality has been substituted for distinctive characteristic exam held by middle schools themselves. Because of the lack of unified quality standard, different middle schools and training institutions have set up their own exam content and item difficulty, which as a result, has heavily increased the economic and psychological burden of parents and children, and worse still, it unevenly created ad enlarge the gap of education condition, resulting in serious unbalanced education development. On this occasion, balanced development of education has been greatly impeded. It's high time that China put monitor of basic education quality on the agenda.

As China's basic education is huge in scale, the wise strategy takes the following. First, set up a pilot experiment of reading literacy monitor net which is consisted of three levels including province, cities and towns. Second, adopt monitor standard and technology provided by scholars and scientific research organization, and make it practicable by city and town educational administrative department. Third, report the achievement result of the average score of cities and towns instead of the individuals. Fourth, provide the provincial and national administrative department with achievement result and its influencing factors as prompt, scientific and accurate data for policy making and adjustment, for curriculum and teaching reform, and for improving children's overall attainment level.

10.7.2 Experiment Localization of International Reading Literacy as Service for Pilot Assessment and Monitor

Multiple linear regression analysis reveals that Chinese test score is one of the two predicators of reading literacy performance. And as is showed above, SCTS is significant with PIRLS2006 reading literacy performance ($r = 0.3$, $p < 0.05$), but

quite asymmetric. The vast majority of high scorers are from more than 90 SCTS, while pupils with average of 85 SCTS are leveled into LSG. Obviously, the reading ability that Guangxi existing Chinese classroom teaching aims to cultivate is different from what PIRLS2006 proposes.

The present Chinese classroom teaching is difficult to satisfy modern children's ad social ever developing need and changes. In China, Chinese classroom teaching is characterized by traditional cultural ideas and former Soviet Union teaching methods in literature classes. Specifically, the characteristics are the following. Reading in Chinese classroom teaching is for learning to read and emotional education; Almost all of the reading text is literary text; reading time is fragmented and insufficient; reading is for the aim of examination, ignorant of children's reading interest and the differences in reading level; in particular, Chinese classroom teaching aims at reaching a kind of standardization and stereotype of emotion, attitudes and evaluation. In a word, the existing Chinese classroom teaching is disadvantageous of cultivating independent thinking and innovative thinking.

It's necessary and urgent to reform the present Chinese classroom teaching and Chinese education to face international challenges. The approach is to set up an experimental school base, in which international reading literacy is tried out and scientific data are provided for reading literacy assessment and monitor and international comparison. Grade 4 pupils, since having been transformed from reading as means to reading as aims, and stepped into the stage of learning through reading, is the most suitable subject of experiment and monitor.

10.7.3 Bring Various Library Services into Full Play, and Make Reading Chance Balance as the Basis of Education Balance

Family BC is significant with family condition ($r = 0.4$, $p < 0.05$), which is the most striking of all the correlation coefficient index. Among the participants whose book collection is below 10, 79 % have intermediate and poor family condition, 91 % mothers and 86 % fathers belong to the lowest OH, i.e. working class. So it's unrealistic to increase their family BC to a large degree. Various libraries have much work to do.

Various libraries should form a networking and make it facility for students to borrow and lend in a quiet, free, and serious environment. Specifically speaking, school libraries can entrust classes and parents to take the responsibility of management, employ class librarians, train parents and students, and make it open at 2-day weekend. Provincial and municipal libraries should stop waiting for readers to enter, and employ volunteers as library liaisons to carry books to schools for exchange and circulation. College libraries should take social responsibility to open to primary and middle school students. What's more, various libraries should get into touch with

charitable publishing houses and social communities, strategically call on them to send and present free books, and also protect their economic interest and cultivate their social reputation. All these measurements aim at balancing children's reading chances and improving balanced development of education.

Appendix

References

Ke Huawei, Zan Yiling, Zhang Jieshu, & You Tingya (2009). PIRLS 2006 Taiwan report of reading literacy in grade 4 (2nd ed).

Lu Xueyi (2010). Hierarchy structure transition of Chinese social classes in 60 years. *Chinese Population Resources and Environment*, 20(7), 1–11.

PIRLS (2006). International Report Exhibit 1.1 distribution of reading achievement [R], 37.

Scientific Research Group of Primary and Middle Schools Students' Academic Achievement in National Institute of Education Sciences (NIES). (2011). Survey Report of Primary Grade 6 Pupils' Academic Achievement. *Educational Research*, (1), 33.

Wright, B. D., & Stenner A. J. (1998). Readability and reading ability. *Paper presented to the Australian Council on Education Research*. http://www.eric.ed.gov/PDFS/ED435979.pdf

Xueyi, L. (2002). Analysis of 10 hierarchies in Modern China. *Learning and Practice, 3*, 55–63.

Zhu Yongxin (2012). Change: Resulting from Reading. *Moral Education China*, (7), 8.

Chapter 11
Extended Mantel-Haenszel Procedure for DIF Detection – A Note on Its Implementation in ACER ConQuest

Xiaoxun Sun

Abstract The latest version of ACER ConQuest is able to report Mantel-Haenszel Statistics to assist differential item functioning (DIF) analysis, which was investigated using log-odds estimators. The purpose of this paper is to focus on the implementation of Mantel-Haenszel Statistics (Mantel and Haenszel, J Natl Cancer Inst 22:719–748, 1959) for both dichotomous and polytomous items in ACER ConQuest.

Keywords Mantel-Haenszel • DIF • ACER ConQuest

11.1 Introduction

The Mantel-Haenszel method, introduced in 1959 by Mantel and Haenszel (1959), is suitable for testing the *null* hypothesis of independence between two dichotomous variables using data from a population subdivided into K groups: it is, therefore, a method for analysing a $2 \times 2 \times K$ contingency table. The Mantel-Haenszel method is used for estimating and testing a common two-factor association parameter in a $2 \times 2 \times K$ table.

X. Sun (✉)
Department of Psychometrics and Methodology, Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell, Melbourne, VIC, Australia
e-mail: xiaoxun.sun@gmail.com

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_11,

Table 11.1 Example of a 2 × 2 contingency table for Mantel-Haenszel statistics computation

	Correct on ith item	Incorrect on ith item	Total
Reference group	A_{ik}	B_{ik}	$A_{ik} + B_{ik}$
Focal group	C_{ik}	D_{ik}	$C_{ik} + D_{ik}$
Total	$A_{ik} + C_{ik}$	$B_{ik} + D_{ik}$	$A_{ik} + B_{ik} + C_{ik} + D_{ik}$

The Mantel-Haenszel method is used to detect DIF. Suppose the interest is in examining whether the dichotomously scored item ***i*** shows DIF for a focal group and a reference group. Typically, the sample is divided into K matched groups based on the raw scores. In ACER ConQuest, the classification of the population is based on population ability estimations EAPs, WLEs, MLEs or PVs, and the estimation type can be specified by the statement `estimate`.

As shown in Table 11.1, $A_{ik}, B_{ik}, C_{ik}, D_{ik}$ represent the number of observations in the kth matched group who belong to the reference group and answered the ith item correctly, who belong to the reference group and answered the ith item incorrectly, who belong to the focal group and answered the ith item correctly and who belong to the focal group and answered the ith item incorrectly, respectively. The K, 2 × 2 tables (one for each matched group) form a 2 × 2 × K table. Under the hypothesis of no DIF, the proportion of correct answers in both reference and focal groups should be the same for all K. The formula for computing Mantel-Haenszel statistics for the ith item is shown as follows:

$$MH\_DIF = -2.35 \times log_e \frac{\sum_K A_{ik} D_{ik} / (A_{ik} + B_{ik} + C_{ik} + D_{ik})}{\sum_K B_{ik} C_{ik} / (A_{ik} + B_{ik} + C_{ik} + D_{ik})}$$

According to the definition, the Mantel-Haenszel statistics is defined to be negative when the item is more difficult for members in the focal group than for the reference group. If there is no DIF, the Mantel-Haenszel statistics is 0. Further, a statistic that follows an approximate χ^2 distribution with $(K - 1)$ degrees of freedom can be computed by as:

$$MH\_CHISQ = \frac{\left(\left|\sum_k A_{ik} - \sum_k E(A_{ik})\right| - 0.5\right)^2}{\sum_k V(A_{ik})}$$

In which, $E(A_{ik})$ and $V(A_{ik})$ are the expected value and variance of A_{ik}, where $E(A_{ik}) = \frac{(A_{ik}+C_{ik})*(A_{ik}+B_{ik})}{(A_{ik}+B_{ik}+C_{ik}+D_{ik})}$, $V(A_{ik}) = \frac{(A_{ik}+C_{ik})*(C_{ik}+D_{ik})*(B_{ik}+D_{ik})*(A_{ik}+B_{ik})}{(A_{ik}+B_{ik}+C_{ik}+D_{ik})^2*(A_{ik}+B_{ik}+C_{ik}+D_{ik}-1)}$ and 0.5 is the Yates' correction for continuity (Yates 1934).

Table 11.2 ETS DIF category

ETS DIF category	P(MH_CHISQ) $\leq$ 0.05	P(MH_CHISQ) $>$ 0.05
|MH_DIF| $\geq$1.5	C = Moderate to large	A = Negligible
1$<$ |MH_DIF| $<$1.5	B = Slight to moderate	A = Negligible
|MH_DIF| $\leq$1	A = Negligible	A = Negligible

In addition to reporting the Mantel-Haenszel statistics, the suggested ETS DIF category is also included in the output of ACER ConQuest. The ETS DIF category (Zwick et al. 1999) is shown in Table 11.2.

11.2 Mantel-Haenszel Method and Its Extension in ACER ConQuest

The implementation of Mantel-Haenszel statistics in ACER ConQuest extends the method to support multiple focal groups and multiple scoring categories.

One possible way to support multiple focal groups is to allow the comparisons between all focal groups and the specified reference group. In the latest version of ACER ConQuest, the Mantel-Haenszel statistics are reported between each focal and reference groups.

In order to support multiple scoring categories, the comparison is made for all pairs of adjacent scoring categories. For example, if a partial credit item has scoring categories 0, 1, and 2; the Mantel-Haenszel statistics is reported between scoring categories 0 and 1, and scoring categories 1 and 2. Please note that this is not the unique way to extend the Mantel-Haenszel statistics to multiple scoring categories, but this approach is consistent with the construction of partial credit model which relies upon the application of the simple logistic model to sequential pairs of response categories.

11.3 Examples and Display of Results

There are two ways in ACER ConQuest to request Mantel-Haenszel procedure. The first way is through `mh` command. The other way is through the `plot` command.

Some examples of requesting Mantel-Haenszel procedure are shown as follows:

Example 1. mh!gins=2,bins=5,estimates=latent,group=gender,reference=F;

In this example, the Mantel-Haenszel procedure is requested for the second item, in which the group is `gender` and the reference group is specified to be `F`. the grouping variable is gender and the reference group is specified to be F. The population has been divided into 5 matched groups based on `PVs`. The ACER ConQuest output of *Example 1* is shown as follows:

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=F
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
================================================================================
item:2 (2)
                Bins                        OBSERVED   EXPECTED
---------------------------------           GROUP      GROUP      CELL       CELL
     Low          High         Mean      N  AVERAGE    AVERAGE    CHISQ      P
--------------------------------------------------------------------------------
      <-       -0.19819     -0.59158    198  1.000      0.970      6.223      0.013
  -0.19819      0.19032      0.00169    202  1.000      0.983      3.508      0.061
   0.19032      0.54286      0.37718    193  1.000      0.988      2.302      0.129
   0.54286      0.96489      0.75860    211  0.995      0.992      0.294      0.588
   0.96489          ->       1.30421    196  1.000      0.995      0.925      0.336
--------------------------------------------------------------------------------
Between ability group fit:   13.252  df=4   p=0.010

================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=M
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:2 (2)
                Bins                        OBSERVED   EXPECTED
---------------------------------           GROUP      GROUP      CELL       CELL
     Low          High         Mean      N  AVERAGE    AVERAGE    CHISQ      P
--------------------------------------------------------------------------------
      <-       -0.19819     -0.58012    204  1.000      0.970      6.339      0.012
  -0.19819      0.19032      0.00935    197  1.000      0.983      3.395      0.065
   0.19032      0.54286      0.38162    207  1.000      0.988      2.458      0.117
   0.54286      0.96489      0.73823    189  1.000      0.992      1.571      0.210
   0.96489          ->       1.32252    203  0.995      0.995      0.004      0.948
--------------------------------------------------------------------------------
Between ability group fit:   13.768  df=4   p=0.008

  /* There are 2 scoring categories */
  Mantel-Haenszel Statistics (between categories 0 and 1) is -0.000; CHISQ=0.001;
df=4; p=1.00000
  Suggested DIF Category is A: Negligible
  The reference group is :F
```

Example 2 presents the Mantel-Haenszel procedure in ACER ConQuest that supports multiple groups.

Example 2. mh! gins=3,bins=3,estimates=wle,group=country,reference=AUS;

In this example, the Mantel-Haenszel procedure is requested for item 3, the grouping variable is country and the reference group is `AUS`. The population has been divided into 3 matched groups based on `WLEs`. The ACER ConQuest's output of *Example 2* is shown as follows:

```
============================================================================
ConQuest: Generalised Item Response Modelling Software
Group: country=AUS
DATA TABLE FOR EXPECTED SCORE
Estimator type is weighted maximum likelihood
============================================================================
item:3 (3)
               Bins                       OBSERVED   EXPECTED
---------------------------------          GROUP      GROUP      CELL       CELL
     Low         High        Mean      N   AVERAGE    AVERAGE    CHISQ      P
----------------------------------------------------------------------------
      <-       0.32868    -0.37637    398  0.349      0.292      6.418      0.011
   0.32868     1.11308     0.88123    201  0.791      0.591     33.161      0.000
   1.11308       ->        1.65531     67  0.970      0.758     16.399      0.000
----------------------------------------------------------------------------
Between ability group fit:   55.979  df=2   p=0.000
============================================================================
ConQuest: Generalised Item Response Modelling Software
Group: country=CHN
DATA TABLE FOR EXPECTED SCORE
Estimator type is weighted maximum likelihood
============================================================Build: Apr 22 2012===
item:3 (3)
               Bins                       OBSERVED   EXPECTED
---------------------------------          GROUP      GROUP      CELL       CELL
     Low         High        Mean      N   AVERAGE    AVERAGE    CHISQ      P
----------------------------------------------------------------------------
      <-       0.32868    -0.08544    248  0.298      0.355      3.475      0.062
   0.32868     1.11308     0.89307    267  0.640      0.594      2.365      0.124
   1.11308       ->        1.84492    152  0.875      0.791      6.436      0.011
----------------------------------------------------------------------------
Between ability group fit:   12.275  df=2   p=0.002
```

```
   /* There are 2 scoring categories */
   Mantel-Haenszel Statistics (between categories 0 and 1) is -0.400; CHISQ=2.929;
df=2; p=0.56975
   Suggested DIF Category is A: Negligible
   The reference group is :AUS
   ================================================================================
   ConQuest: Generalised Item Response Modelling Software
   Group: country=USA
   DATA TABLE FOR EXPECTED SCORE
   Estimator type is weighted maximum likelihood
   ============================================================Build: Apr 22 2012===
   item:3 (3)
                Bins                        OBSERVED   EXPECTED
   ---------------------------------          GROUP      GROUP       CELL        CELL
        Low          High         Mean       N  AVERAGE   AVERAGE     CHISQ        P
   --------------------------------------------------------------------------------
         <-       0.32868      0.10757     168  0.119      0.400      55.375      0.000
    0.32868       1.11308      0.92625     218  0.518      0.602       6.400      0.011
    1.11308         ->         2.07120     281  0.786      0.826       3.105      0.078
   --------------------------------------------------------------------------------
   Between ability group fit:   64.880  df=2   p=0.000

   /* There are 2 scoring categories */
   Mantel-Haenszel Statistics (between categories 0 and 1) is -1.183; CHISQ=_BIG_;
df=2; p=0.00035
   Suggested DIF Category is B: Slight to moderate
   The reference group is :AUS
```

Example 3 demonstrates the Mantel-Haenszel procedure in ACER ConQuest supports the multiple scoring categories as well.

Example 3. mh! gins=1,bins=5,estimates=latent,group=gender,reference=F;

In this example, the Mantel-Haenszel procedure is requested for item 1 with three scoring categories, the grouping variable is gender and the reference group is specified to be `F`. The population has been divided into 5 matched groups based on `PVs`. The ACER ConQuest output of the *Example 3* is shown as follows:

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=F
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
================================================================================
item:1 (1)
            Bins                        OBSERVED   EXPECTED
-----------------------------------      GROUP      GROUP       CELL       CELL
     Low         High         Mean     N  AVERAGE    AVERAGE     CHISQ      P
--------------------------------------------------------------------------------
      <-       0.16097    -0.24979    182  0.599      0.437      10.948     0.001
 0.16097       0.53659     0.37551    210  0.814      0.774       0.543     0.461
 0.53659       0.89430     0.72553    204  1.103      1.003       3.065     0.080
 0.89430       1.28020     1.07357    206  1.112      1.230       4.614     0.032
 1.28020         ->        1.62958    198  1.343      1.535      15.894     0.000
--------------------------------------------------------------------------------
Between ability group fit:   35.062  df=4   p=0.000
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=M
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:1 (1)
            Bins                        OBSERVED   EXPECTED
-----------------------------------      GROUP      GROUP       CELL       CELL
     Low         High         Mean     N  AVERAGE    AVERAGE     CHISQ      P
--------------------------------------------------------------------------------
      <-       0.16097    -0.24694    219  0.603      0.438      13.569     0.000
 0.16097       0.53659     0.36543    190  0.932      0.768       8.150     0.004
 0.53659       0.89430     0.74078    197  1.020      1.013       0.016     0.901
 0.89430       1.28020     1.07338    195  1.138      1.230       2.608     0.106
 1.28020         ->        1.67917    199  1.352      1.557      19.101     0.000
--------------------------------------------------------------------------------
Between ability group fit:   43.444  df=4   p=0.000
```

```
/* There are 3 scoring categories */
Mantel-Haenszel Statistics (between categories 0 and 1) is 0.017; CHISQ=0.001; df=4;
p=1.00000
Suggested DIF Category is A: Negligible
The reference group is :F

Mantel-Haenszel Statistics (between categories 1 and 2) is 0.039; CHISQ=0.042; df=4;
p=0.99978
Suggested DIF Category is A: Negligible
The reference group is :F
```

Example 4 shows the other way to request Mantel-Haenszel statistics, which is to use by `plot` command.

Example 4. plot expected! gins=3,bins=5,table=yes,estimates=latent,group=gender,mh=F;

In this example, the Mantel-Haenszel procedure is requested from the plot command, and it calculates the Mantel-Haenszel statistics for item 3, the grouping variable is gender and the reference group is specified to be `F`. The population has been divided into 5 matched groups based on `PVs`. The ACER ConQuest output of E*xample 4* is as follows:

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=F
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
================================================================================
item:3 (3)
                Bins                        OBSERVED   EXPECTED
---------------------------------           GROUP      GROUP      CELL       CELL
     Low         High        Mean      N    AVERAGE    AVERAGE    CHISQ      P
--------------------------------------------------------------------------------
      <-       0.12583    -0.30123    208   0.298      0.307      0.079      0.779
 0.12583       0.50174     0.34824    184   0.484      0.459      0.452      0.501
 0.50174       0.85885     0.68975    206   0.583      0.544      1.222      0.269
 0.85885       1.30104     1.06090    202   0.673      0.634      1.360      0.243
 1.30104         ->        1.70622    200   0.730      0.767      1.566      0.211
--------------------------------------------------------------------------------
Between ability group fit:   4.679  df=4   p=0.322


================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=M
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:3 (3)
                Bins                        OBSERVED   EXPECTED
---------------------------------           GROUP      GROUP      CELL       CELL
     Low         High        Mean      N    AVERAGE    AVERAGE    CHISQ      P
--------------------------------------------------------------------------------
      <-       0.12583    -0.29723    193   0.295      0.308      0.143      0.705
 0.12583       0.50174     0.32425    216   0.463      0.453      0.086      0.770
 0.50174       0.85885     0.69555    196   0.536      0.546      0.077      0.781
 0.85885       1.30104     1.06600    196   0.633      0.635      0.004      0.948
 1.30104         ->        1.68689    199   0.784      0.764      0.442      0.506
--------------------------------------------------------------------------------
Between ability group fit:   0.752  df=4   p=0.945

   /* There are 2 scoring categories */
   Mantel-Haenszel Statistics (between categories 0 and 1) is -0.054; CHISQ=0.071;
df=4; p=0.99939
   Suggested DIF Category is A: Negligible
   The reference group is :F
```

By using the `plot` command, the item plot across the reference and focal groups will also be shown on the screen.

11.4 Discussions

Another widely used DIF detection approach is the standardized item difficulty difference method. The method applied χ^2 test of the standardized item difficulty difference between the reference and focal groups. In this method, the standardized item difficulty difference and the significance level of the χ^2 test can be used as the threshold to assist IF detection. A pair of confidence interval lines can be drawn to show DIF items graphically. The significance level of the χ^2 test becomes very sensitive when the sample size is large, and there is no consensus on how to choose the threshold of standardized item difficulty difference. This approach does not work for partial credit items, but work for multiple groups. The Mantel-Haenszel procedure implemented in ACER ConQuest can handle both multiple scoring categories and multiple groups.

11.5 Conclusion

In this note, the implementation of Mantel-Haenszel method in ACER ConQuest was discussed, and some running examples and sample syntax have been given for instructions.

References

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society, 1*(2), 217–235.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

Chapter 12
A Research on the Effectiveness of DynEd Computer-Assisted English Language Learning – Taking Ningbo Polytechnic as an Example

Jingru Huang and Baixiang Wu

Abstract The integration of multimedia-aided English language teaching and learning is one of the trends of abroad and domestic research. DynEd CALL (computer-assisted English Language learning) is a new answer to the challenge of the integration research. This paper explores the DynEd CALL model based on public English teaching reform of Ningbo Polytechnic that has been done for 3 years, which aims to find the effective model of language learning and teaching. The purpose of this study is to investigate the effectiveness of DynEd CALL model. The data were collected through DynEd Records Manager, on over 1,800 non-English majors from eight institutes of Ningbo Polytechnic, implemented for 1 year. The data included each student's two Placement Tests, two Speaking Tests (both were done at the beginning and the end of learning DynEd), and detailed information of learning process about the courseware such as studying days, total study hours, study frequency, study score, functional buttons usage and speech recognition. The data also contained a questionnaire followed by 1 year's DynEd study. Pre-and post-comparative analysis of two Placement Tests and Speaking Tests show that the differences are nearly 0.5 and 0.3 respectively, which means the award-winning, multimedia content keeps students on task and engaged. And the questionnaire indicates students have adopted the concept that language is a skill, not knowledge. DynEd CALL can significantly promote students English learning proficiency, esp. on listening and speaking skills. The improvement on reading and writing skills, however, are not so dramatic comparatively, which are the perspectives we should enhance in the future teaching and research.

J. Huang (✉)
Department of Basic Education, Ningbo Polytechnic, Ningbo, Zhejiang, China
e-mail: sallyhuang04@126.com

B. Wu
Department of International Business, Ningbo Polytechnic, Ningbo, Zhejiang, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_12,

12.1 Introduction

Language learning is an integrated process that involves with teaching, learning and testing. The three parts are often practiced independently, but it will not be very effective without concerning the other two parts. Heaton (2000) commented that "testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other. Tests may be constructed primarily as devices to reinforce learning and to motivate the student or primarily as a means of assessing the student's performance in the language" (Heaton 2000, p.6).

It is a tradition that testing, teaching and learning are practiced independently in most Chinese university English teaching classrooms. There are two kinds of English tests often taken in their college years for Chinese full-time undergraduates who are majoring in any disciplines except English which are college English tests at the end of each semester administered by their own colleges/universities and CET-4/CET-6 (College English Test Band 4/6) administered by National College English Testing Committee on behalf of the Higher Education Department, Ministry of Education, the People's Republic of China. The college English test of each semester is based on what students have learned in college English class. It aims to check on student English learning progress, provide feedback for teachers and students, and facilitate subsequent teaching and learning. CET-4/CET-6 measures the English proficiency of undergraduates in accordance with the National College English Teaching Syllabus (NCETS), which exerts tremendous impact on college/university teaching and learning of English in China and affects the huge numbers of stakeholders (see Cheng 2008 for a full review). The achievement of CET-4/CET-6 is obvious. It has served as a good benchmark for English language teaching and learning in the past two decades in China (Li 2002; Gu 2005). There is also drawback as the benchmark is very much blamed as a reason of the inefficiency of Chinese English language learners in communication (e.g. Han et al. 2004; Gu and Liu 2005).

We do agree on Heaton's statement, however, that testing, teaching and learning should be an integrated system and should not be treated independently (Heaton 2000). This paper fully emphasizes on the effectiveness of DynEd CALL, an integrated system on language teaching and learning model and its outstanding test system.

12.2 Background

Vocational higher education is a very important part of higher education in China. There are 2.99 million students enrolled by the higher vocational colleges nationwide and make the total number of vocational college students reach as many as 96 million, almost the same number as the regular university students in 2007 (Ma 2012).

According to the Fundamental Requirements in English Courses of Higher Vocational Education formulated by the Higher Educational Department of the Ministry of Education of China, that the job for vocational college is to cultivate talents of practice-oriented and craft-oriented and the purpose of English learning is not for conducting scientific research or academic exchanges, but for improving the basic humanities accomplishment or the ability to engage in foreign-related business (Ministry of Education 2006). Here the guiding principle is practicality and sufficiency. Listening and speaking training should have the priority to foster students the basic English ability in the workplace environment.

For a long time, the teaching approach and test of higher vocational English have been influenced by the traditional teaching concept. Summative evaluation based on semester final test paper plays an important role in the measurement of students' semester English learning effect. The result of this evaluation leads to students pay more attention to test scores rather than the process of learning. Traditional teacher-centered approach occupies most vocational college English classroom (Zhang 2011a). And this evaluation system neglects students' ability, interest and diversity of learning styles and it pays too much emphasis on the role of selection and elimination. The direct result of this single mode evaluation leads to the fact that both teaching and learning are for the tests only. It limits the teacher's curriculum content and represses the students' independent learning ability (Guo 2003).

As for the English semester final test, many researchers express their deep concern about its deficiency and negative backwash to the teaching and learning. The formats of English semester final test are deeply influenced by CET-4. The typical feature of the test is that multiple choices occupy much proportion of the test and oral test usually is not included (Zhou 2002; Yang 2008). As we all know, CET-4/CET-6 is a criterion-related norm-referenced test (National College English Testing Committee 2006). More than millions of test-takers take part in CET-4/CET-6 each year. The large-scale and high-stakes of the test determines it has to adopt standardized language test using more objective multiple choices. Multiple choices has some disadvantages such as encouraging guessing, testing the students' ability of language knowledge recognition rather than the ability of language usage, etc. (Hughes 2000). Furthermore, limited by their teaching concept and knowledge scope, common English teachers sometimes can't come up with a satisfactory test paper. "Measurement error occur in test scores, evidence for validity is insufficient, some items are too difficult or too easy, or fail to effectively discriminate between weak and strong students in the ability being tested" (Gu 2010, p.119). Hence, English final tests in the frame of CET-4 format can't effectively feedback the teaching and learning (Zhang 2011b; Chen 2011; Yuan 2012).

As Hughes (2003) said, "The proper relationship between teaching and testing is surely that of partnership" (Hughes 2003, p.6). Can we find an effective way to assist students to improve their listening and speaking ability? Is there a test simple, but inclusive and can fast and effectively reflect students' progress in learning? Can the test give positive backwash to teachers and students?

In order to solve these problems, Ningbo Polytechnic (NP) has carried out its public English teaching reform for more than 3 years. The teaching reform experiment is carried out with an integrated language test and learning system DynEd CALL to assist college English teaching and learning. In the fall of 2008, altogether 200 students from five majors (Tourism, Marketing, Computer, Architecture and International trade) participated the pilot test. After 1 year's learning, most of these students have made some remarkable progress in English, especially in listening and speaking. Then, from 2009 until now, freshmen of non-English majors in NP use DynEd courseware for their college English learning. This research focuses on the following questions:

1. Can DynEd CALL improve students' English proficiency, especially in listening and speaking?
2. What's the students' DynEd learning process during one school year?
3. What's the backwash of tests to learning and teaching?

The hypothesis behind the research questions is that DynEd CALL is effective to help students break dumb English and confidently make basic oral English communication especially for learners from polytechnics whose English proficiency is comparatively at a low level.

12.3 Related Theories

DynEd International, Inc. was founded in 1987 in America by a team of language teachers, engineers, and artists. And DynEd created the world's first interactive multimedia language learning CD-ROM in 1988 and received a U.S. patent 3 years later. DynEd's courses cover all proficiency levels from kids to adults. The courses are used in more than 50 countries including France, Korea, Malaysia, Mongolia, Turkey, and China.[1]

12.3.1 Evolving Role of CALL

Computer Assisted Language Learning (CALL) is an emerging force in language education. Well-designed multimedia lessons can coordinate visual, auditory and contextual input, but a book or language lab cannot do in the same way.

Research shows that listening and reading skills use different pathways within the brain. According to Deacon, a neuroscientist, the speaker and listener have to generate associated words rapidly when they speak or listen. They should be to

[1] The information is retrieved from DynEd website in Aug. 2, 2012. (http://www.dyned.com/us/about/)

avoid letting earlier associations interfere the process of generation automatically. Thus the cognitive process must be simple but rapid enough (Deacon 1997). That is to say, our brain's cerebellum is involved in auditory processing and the auditory pathway is much faster. If learners are exposed to auditory input with text support, this process must set up auditory and visual input. And this competing input makes it more difficult to decode the incoming speech in the necessary auditory processing speed. As another neuroscientist, Richard Restack says, "Competition between sensory channels can also prove disruptive" (Restak 1994, p.79).

We have fully realized listening is the key skill in language learning, but our English classes still mainly depends on text-based and reading activities as our primary language materials. We are accustomed to this form and enjoy it because text gives us time for conscious analysis. Though it may be comfortable, research indicates it isn't effective because it eliminates the temporal tension. When we are learning English, we need temporal tension to form chunking ability. The appropriate amount of temporal tension can lead to attention, efficient practice, and language automaticity. Traditional language education is knowledge-based and it focuses on event memory. Teachers are knowledge givers and learners memorize facts, rules and definitions. Even if students may have a large vocabulary and a good understanding of grammar, they may not be able to communicate effectively because traditional language education fails to develop automaticity. According to Knowles (2004), automaticity means that learner can understand the speech without the need refer of the text. Grammar, syntax and vocabulary all can be put aside temporarily. Using CALL, the visual and auditory input is delivered and the learners react with the presentation for a sense of automaticity (Knowles 2004). The visual display of an icon can activate many areas of the brain. When we recognize a familiar object or icon, our brain can utilize the knowledge, concepts, and associations about that similar object to decode the meaning of a string of sounds. And this iconic presentation can help learners bootstrap the language learning process. According to Knowles, 4-skills path of language learning should be progress from listening to speaking, to reading and then to writing and CALL lessons can play an important part in providing this kind of practice, especially the repetitive practice. Repetitive practice is at the heart of skills development (Knowles 2004) .

12.3.2 RHR – A Brain-Based Language Learning

12.3.2.1 Hierarchical Structure of Memories

The automaticity involves the learning theory of RHR (Recursive Hierarchical Recognition). The hierarchical structure of memories and concepts is a key feature in RHR. From the neurosciences, we know that Episodic Memory is responsible for explicit memory (event and fact learning), and it needs learning with awareness.

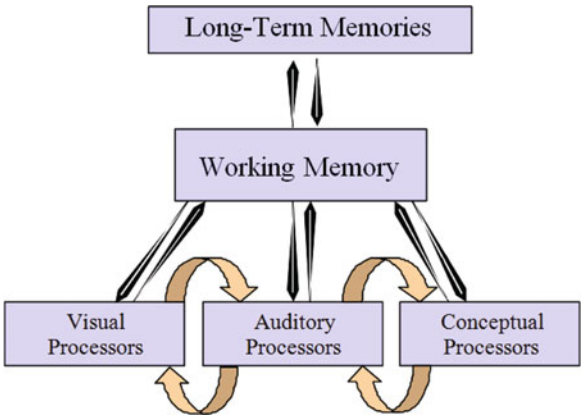


Fig. 12.1 The way of language processing (*Source*: Knowles 2008)

Procedural Memory is responsible for implicit memory (skill learning), and it needs learning without awareness (Restak 1994). A skill is acquired by involving the activation of an automatic sequence of actions through repetition and/or practice over a suitable period of time. Language is a skill, not knowledge. Learning language is just like learning swimming, or playing a musical instrument. All these activities need repetitive practice until they are stored in procedural Memory. Once the skills are formed, they become automaticity and can't be forgotten.

As the famous neuroscientist, Donald Hebb said as early as 1949: "Neurons that fire together, wire together... that repeated excitations of a sequence of neurons modifies the synaptic connections between those neurons" (Hebb 1949, p.62). This is the basis for Hebbian learning. From the neural sciences, we know that "multimodal activities can strongly enhance the creation of new or strengthened synaptic connections and these synaptic connections are the stuff of new memories, especially procedural memories" (Knowles 2008). Knowles presented an oversimplified diagram (Fig. 12.1) to show how various processors in the brain communicate with each other and the working memory. The way of processing language is just like Fig. 12.1 shows: Language processing requires many neural systems to interact and these processors such as auditory, visual and conceptual work in parallel and interacts with the working memory and long-term (LT) memory upward or downward in order to piece together and interpret language and sensory input.

Some neuroscientists believe long term (LT) memory, visual information, and conceptual processors work together to help decode and fill-in comprehension gaps. Our brain has natural ability to make sense of things and fill-in details or patterns to fit one's expectations. This conceptual structuring is done through millions of tiny cortical columns in the brain's neocortex. Each tiny cortical column processes a specific type of sensory input. When many different columns are switched again and again they will form a network and activate as a whole. It will very much increase the reaction speed of language input and process the language in chunk instead of vocabulary and grammatical information (Knowles 2008).

12.3.2.2 Chunking Ability and Concepts

As the cognitive scientist Steven Pinker points out: "Phonological short-term memory lasts between 1 and 5 s and can hold from four to seven chunks. Short-term memory is measured in chunks rather than sounds because each item can be a label that points to a much bigger information structure in long-term memory, such as the content of a phrase or sentence."(Pinker 1997, p.89) We feel pressure when we try to hold auditory information in limited memory buffers and this creates temporal tension. Appropriate temporal tension can engage and motivate the learner effective learning.

In RHR, language chunks are built around concepts and these concepts express elements of information or language functions. In RHR, the goal is for students to master the framework of the language. The grammar and syntax related to the concepts is the framework of the language, just like the trunk and branches of a tree. Vocabulary and expressions are like the leaves of a tree. The trunk and branches of the language framework hold the language together. And this framework helps students learn and remember vocabulary. It provides the patterns for chunking. RHR develops the language chunking ability by using the fact that language is a system of patterns, and the brain is very good at recognizing and using pattern. The ability to automaticity and process language chunks is the key to language frequency.

12.3.3 Blended-Model

Blended-model refers to a blend of computer and classroom. Computer lessons provide the language framework and lots of opportunities to understand and practice these key patterns. In the RHR approach, the key patterns of English are carefully introduced so that the brain learns to recognize and use them. The learning sequence is: (1) familiarization (2) recognition (3) comprehension (4) master and (5) automaticity. Our brain accomplishes this sequence through the use of multi-modal input, which is made possible by computers. Multimodal means involves multiple parts of the brain, such as sight, sound, and physical action. The input and practice will work in a recursive, circular manner and will eventually form pattern-recognizing subroutines (Knowles 2008). In the classroom, these language models are personalized and extended by the carefully designed activities. The classroom teaching provides the human element and language extension. This combination allows learners to approach language study much more effectively. In this blended model, the teacher's role is changed. We are not knowledge givers just like in the traditional knowledge-based approach. Now we are coaches and communication facilitators. This blended approach is different from other computer-assisted language learning (CALL) approaches which may or may not involve teachers and classroom support.

12.4 Method

12.4.1 Participants

A total of 1,888 non-English major freshmen from Ningbo Polytechnic were enrolled in this DynEd study experiment that lasted more than 30 teaching weeks in two semesters scheduled from Oct, 2009 to June, 2010. These 1,888 non-English major freshmen came from 53 classes (19 liberal arts, 34 science classes) in eight different institutes.

12.4.2 Teaching Design

College English is a compulsory course, weighing ten credits in our credits requirements. There were four teaching periods per week, two periods were lab-teaching and the other two were classroom teaching. Besides these four teaching periods, students were required to do autonomous online courseware learning to finish the assignments in the lab. The total duration of online courseware learning time was at least 4 h per week. The learning frequency was required no less than 3 days per week.

12.4.3 Procedure

12.4.3.1 Teacher Training

All the college English teachers partook in the DynEd teacher training program given by the DynEd experienced trainers in the summer vocation of 2008. All these teachers transformed teaching concept and studied DynEd courseware seriously. Most teachers' online studying hours were more than 50 h and we got familiar with some of the DynEd courseware, for example, *New Dynamic English*, *First English* and *Teacher Training*. All these efforts set a solid foundation for the coming large-scale DynEd teaching.

12.4.3.2 First Placement Test (PT)

To place these freshmen at the appropriate starting point in DynEd's series of English language courses, they were asked to take a placement test shortly after their enrollment in September. Before giving the Placement Test, teachers gave students a placement test orientation by providing basic instructions and examples of each type of question. There were several types of questions in the Placement Test, including multiple-choice items that test vocabulary and grammar, listening

Table 12.1 DynEd placement level and relating appropriate courses

DynEd placement level	Appropriate courses
0.0–0.2 Beginner ~	New Dynamic English Module 1; First English
0.5–0.7 (TOEIC 250–450)	New Dynamic English Mod 2; FE Units 5–8; English For Success Units 1–4; The Lost Secret
1.0–1.2 (TOEIC 400–550)	NDE Mod 3; The Lost Secret, EFS Units 5–10
1.5 (TOEIC 500–650)	NDE Mod 4; The Lost Secret, EFS Units 5–10
2.0 (TOEIC 600–750)	NDE Modules 5&6, Functioning in Business, Dynamic Business English 1,2,3,4; Hospitality English
2.5 (TOEIC 650–800) – TOEFL 540	NDE Mod 7; FIB; DBE 3,4,5,6; Dialogue; Test Mountain
3.0 (TOEIC 800–950) – TOEFL 630	NDE Mod 8; Dialogue; Test Mountain; Advanced Listening
3.5–5.0 (above DynEd's scope)	Advanced Listening; Test Mountain

Source: DynEd placement test, http://www.dyned.com

comprehension, sentence construction, and sentence ordering. Then students had the practice test several times before they began the placement test.

There are two parts in Placement Test. Part 1 places students into two groups. Students who are at DynEd Levels 0.0–1.2 will study part 1 and students who are within level 1.2–3.5 will go to part 2. Students with higher level are considered do not need to study DynEd's courses.

The questions in Placement Test are computer adaptive with variable length. If students performed well it ramp up more quickly to higher-level. Therefore, each student is answering a different test with questions randomly selected from a database.

Based on their Placement Level, students could be placed into appropriate classes and took appropriate courses according to Table 12.1:

12.4.3.3 Daily Study

Learning Content

After first placement test, 53 classes were divided into two groups with classes of advanced level whose average PT level was 0.5 or higher, and classes of common level whose average PT level was lower than 0.5. The two groups are following different study paths with the help of teachers. Advanced classes and common classes learned DynEd courseware from different starting point and at different learning rate. After one school year's study, all the students covered the first three modules of NDE (New Dynamic English) and all the 8 units of FE (First English). Students in advanced class learned module 4 of NDE and the Lost Secret.

Learning Method of Courseware

In the lab-teaching and autonomous online courseware study, content is introduced in a suitable context through multimedia-based listening. In the core course

New Dynamic English, there are two lesson types: presentation lessons and support lessons.

Presentation lessons are the most important part of the course, which introduce and develop key language models. Support lessons follow presentation lessons and include lessons such as question practice, focuses exercises and speech recognition and video interaction.

Students studied presentation lessons in 5-step sequence: Preview, Comprehension, Focused Practice, Mastery and Automaticity. Focused Practice involved production. Students used Voice Record feature to practice saying key sentences. This was where the chunking ability developed. Repeating longer sentences and phrases without text support required concentration and temporal tension. After recording a sentence, students should compare their recording with the model of the course. This comparison helped the brain to improve the students' pronunciation and to set up new neuron connections. This kind of comparison was a feedback loop and it provided information to the brain about to how to adjust performance. Mastery of the language models slowly became automatic in the following reviews. Students went through a lesson several times on different days, from limited comprehension to full comprehension, from summarizing parts of the lesson to making oral presentations until to automaticity.

Student could use Completion Percentage, a function in the courseware, to measure how active he was in a lesson. Each lesson has a target number of learning steps that students should reach for that lesson. We encouraged students to reach 100 % completion.

DynEd's patented shuffler helps students to optimize the cognitive load in each activity. As learner gains proficiency, the course opens more content and the depths of the material increases. For learners having difficulty, the material recycles more often for helping students' comprehension. The high shuffler level is 3.0. And we encouraged students to reach 3.0 and remained at 3.0 shuffler level of each lesson to ensure study effect.

Examination Mode

In each module of NDE, there are three mastery tests of each unit and one module mastery test. Mastery test is designed to evaluate student progress and confirm mastery of the language skills developed in each course. Each test consists of 20–50 test items and it is computer-assisted achievement test. Test items are randomly selected from a database of test items. And the mastery test scores are automatically recorded in the Study Records. The best score is 100. If the score is lower than 85, it means this student fails to pass this mastery test and he needs more study and another chance to do this mastery test again.

Students in the experiment project do not need to attend other semester English final test. In the project a formative assessment was used to measure students' college English learning effect. The formative assessment consisted of three parts with 40 % on students learning process, 40 % on the content proportion they have

Table 12.2 Types of test questions in the Speaking Test (0.0–1.0)

Component	Items	Item type	Focus on	Maximum score
Part 1	8	Sentence reading	Entire sentence	32
Part 2	4	Reading & pronunciation	Key phoneme contrasts	16
Part 3	5	Comprehension & structure	Comprehension and/or grammar	20
Part 4	8	Sentence repetition (no text support)	Sentence repetition the ability to automatically process and chunk language in working memory skill	32

learned and another 20 % on students' attendance of classes, oral presentation and written assignment. The learning process was made up of every week's learning time, frequency, completion percentage of assignment and NDE's study score in 2 weeks. The score of content proportion was from various Mastery tests of the semester.

12.4.3.4 First Speaking Test (ST)

After one semester's learning, students accumulated some input of language patterns and formed some chunking ability. Then, at the beginning of the 2nd semester in February, we began to use Speaking Test to check students' speaking ability.

The Speaking Test helps determine oral fluency level. It uses advanced speech recognition technology and it must be given in a controlled environment where external or background noise is minimal, and where a suitable microphone is used. Like Placement Test, teachers must give students a brief orientation about the test and students should take the Practice Test several times to familiarize themselves with the test format and ensure the audio and microphones were working properly. Once the Speaking Test began, students should not exit the test until it was completed and their score was recorded. There are two levels in the Speaking Test with the lower-level Speaking Test scores between 0.0 and 1.0 and higher-level scores between 1.0 and 2.7 plus. Lower-level Speaking Test (0.0–1.0) has four parts as recorded in Table 12.2: sentence reading, reading & pronunciation, comprehension & structure, and sentence repetition. Each part is scored separately, and the total score is the sum of the part scores.

There are 25 items altogether in the speaking test which generally takes 4–8 min to finish. Maximum score of each item is 4. If recognized on the first attempt, the score will range from 2 to 4, depending on the confidence level as determined by the Speech Recognizer. If recognized on the second attempt, the score will be 2, regardless of confidence level.

Higher-level Speaking Test has two parts (See Table 12.3). As the test progresses, test items become longer and more complex for measuring student's chunking skill. The test will stop if too many items are missed in this level.

Table 12.3 Types of test questions in the Speaking Test (1.0 to 2.7+)

Component	Items	Item type	Focus on	Maximum score
Part 1	5	Sentence reading	Entire sentence	25
Part 2	8–20 (depending on performance)	Sentence repetition (no text support)	Sentence repetition, the ability to automatically process and chunk language in working memory skill	100

Table 12.4 Oral production level for the Speaking Test (0.0 to 1.2+)

Test score	Oral Prod Level
0–27	Beginner or mistest
28–37	0.2
38–51	0.5
52–67	0.7
68–81	1.0
82–100	1.2 or higher

Source: DynEd teacher guides, http://www.dyned.com

Table 12.5 Oral fluency level for the Speaking Test (1.0 to 2.7+)

Test score	Fluency level
0–28	Beginner or mistest
29–36	1.0
37–44	1.2
45–55	1.5
56–59	1.7
60–67	2.0
68–72	2.2
73–84	2.5
85–100	2.7 or higher

Source: DynEd teacher guides, http://www.dyned.com

The maximum score for a recognized item in this part is 5 instead of 4. The maximum score for an item recognized on the 2nd time is 3.

A summary of scores and the estimated Oral Production Level for the Speaking Test (0.0 to 1.2+) is reported in Table 12.4.

Table 12.5 presents a summary of scores and the estimated oral Fluency Level for the Speaking Test (1.0 to 2.7+).

Students are rated into five levels which is called DynEd Placement Levels from 0 to 5. If the Oral Production Level (0.0–1.0) or Fluency Level (1.0–2.7) is higher than the student's Placement Level, then the Placement Level is going to be adjusted upward. If the Speaking Test Level is lower than the Placement Level, the Placement Level is unchanged.

12.4.3.5 Second Placement Test (PT) and 2nd Speaking Test (ST)

After nearly 1 year's learning, at the end of June, we retested students the Placement Test and Speaking Test. Comparing students entry and exit scores of PT and ST, the resulting scores could be fairly good indicators of progress.

12.4.3.6 Questionnaire

After 1 year's DynEd learning, students were asked to finish a simple questionnaire, which had 18 items about Dyned study. We randomly selected 200 students from eight different institutes. Totally 200 questionnaires were sent out and 189 valid ones were retrieved with the effective rate of 94.5 %. Response format was 7-point Likert-type scale (strongly agree = 7, strongly disagree = 1)(refer to Table 12.8 of questionnaire).

12.4.4 Data Acquisition and Analysis

This study uses a variety of methods for data collection, includes: (1) students' learning process, such as studying hours, studying frequency, studying scores; (2) 1st and 2nd Placement test; and (3) 1st and 2nd Speaking test and questionnaire. The data of students' learning process were collected from DynEd's Records Manager, which monitors study activities for class and individual students. We mainly compare the differences between 1st and 2nd Placement test and 1st and 2nd Speaking test and try to find the positive and influential elements of studying. Descriptive statistics are used for the questionnaire results.

12.5 Results and Analysis

After 1 year's studying, students have made some remarkable progress in English proficiency. We can verify the progress in the following ways to answer the three research questions: (1) The overall improvement of 1st and 2nd PT and ST; (2) Students' learning process; and (3) Comparison of students' progress based on students' courseware studying data.

12.5.1 The Overall Improvement of 1st and 2nd PT and ST

The first research question pertained to the learning effects of students' 1 year studying. From students' 1st and 2nd Placement Test, we can see the overall average PT level increased from the 1st of 0.5 to the 2nd of 1.0, uprating two levels. The other important piece of information we can obtain from Fig. 12.2 is that

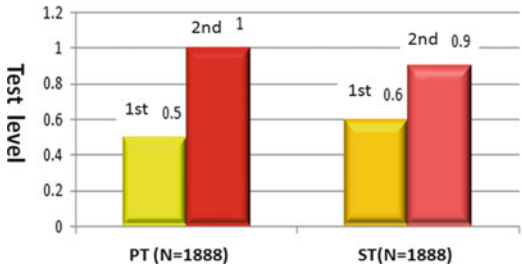


Fig. 12.2 Comparison of 1st and 2nd average levels of PT and ST

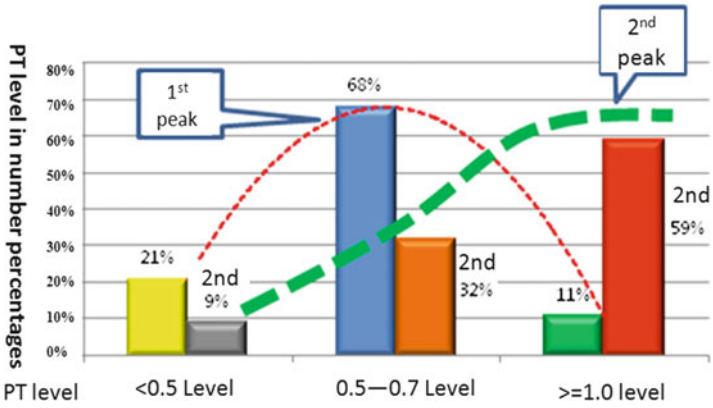


Fig. 12.3 Comparison of 1st and 2nd PT level in number percentages (N = 1,888)

students have made 0.3 growth rate in Speaking Test from 0.6 to 0.9 in short interval of 3 months. This indicates that the first semester's language input has accumulated the oral skills. Students have formed some chunking ability. That is why they could make great progress in the second semester.

We can analyze these variables in details from Figs. 12.3 and 12.4. In Fig. 12.3, we compare the participants' 1st and 2nd PT level in number percentage. In the 1st PT, 68 % of the students were in the level of 0.5–0.7, but this percentage decreased to 32 % in the 2nd PT; meanwhile, the percentage of 1.0 or higher has increased from 11 % in the 1st PT to 59 % in the 2nd PT. And the percentage of the lower level of less than 0.5 decreased from 21 to 9 %. All these figures show the obvious improvement in students PT level.

The PT improvement can be analyzed in another way – the growth rate. We can draw some conclusion from Fig. 12.4 that shows 79 % students have made various degrees of improvement including 31 % students' growth rate in 0.1–0.4, 26 % on 0.5 and 9 % students more than 1.0, which is a marvelous growing.

The Speaking Test Level can be analyzed in the same way as PT level. From Fig. 12.5, we know 56 % were in the level of 0.5–0.7 in the 1st ST but this percentage dropped to 26 % in the 2nd ST. And there are 52 % students achieved 1.0 or higher with 7 % students got to 2.0, which is considered as exceptionally high

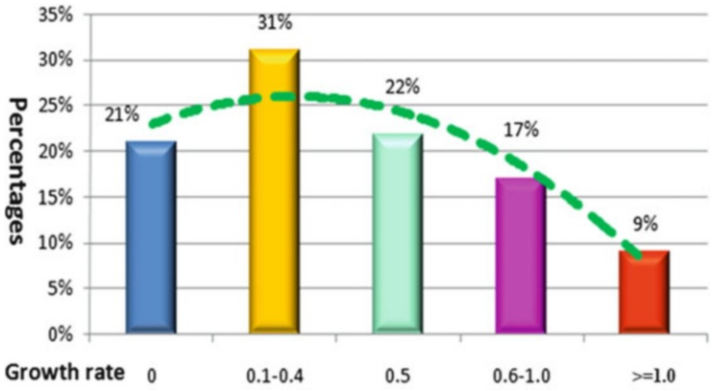


Fig. 12.4 Growth rate of 2nd PT Level in percentages (general average growth rate: 0.5)

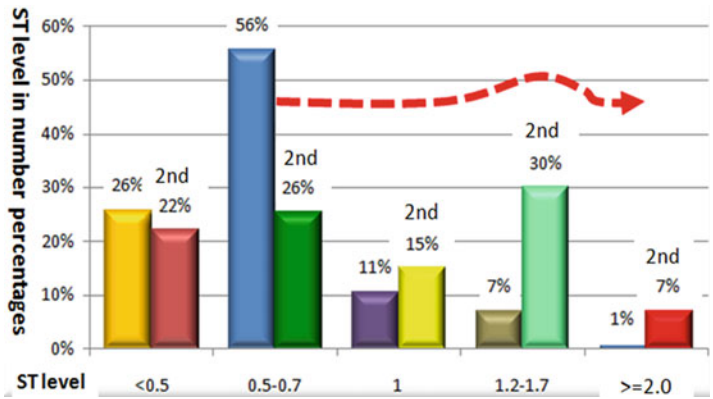


Fig. 12.5 Comparison of 1st and 2nd ST level in number percentages (N = 1,888)

level for these students. Figure 12.6 shows that about 63 % of the participants made the growth rate between 0.1 and 1.0. All of these figures made us conclude that students have made astonishing progress in ST. Refer to Figs. 12.5 and 12.6 for the figure of growth.

Improvements of PT level and ST level provide a good indication that most participants have made some progress in English proficiency, esp. in listening and speaking after 1 year's learning.

12.5.2 Students Learning Process

According to the DynEd Records Manager, the average DynEd teaching weeks of 2009 were 30. Among 53 classes, 55 % of which did 31–34 weeks of DynEd

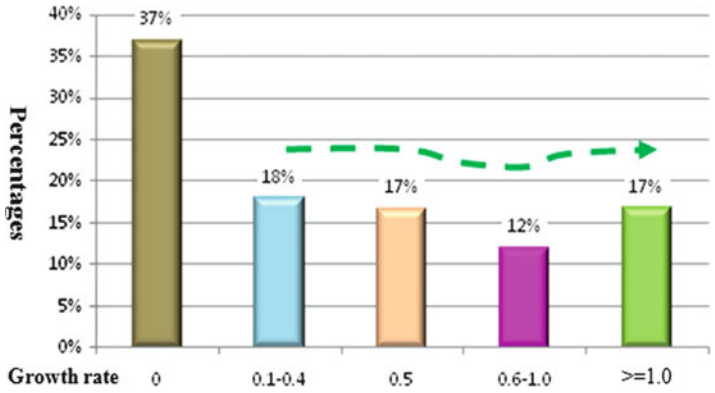


Fig. 12.6 Growth rate of 2nd ST level in percentages

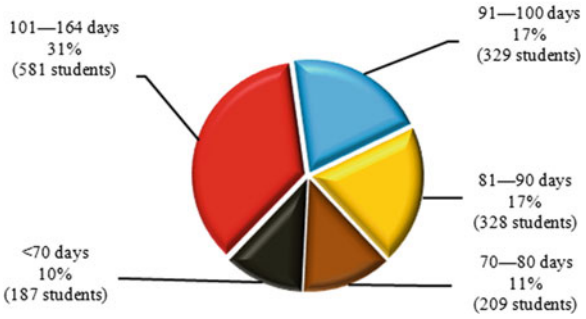


Fig. 12.7 Proportion of students study days ($N = 1{,}888$)

teaching with the other 45 % classes (most were science classes) did 26–30 weeks of teaching. That is to say, most science classes did not reach the average teaching week because of professional training and the students studying time was not enough. In some way, the insufficient practice affected these students progress.

In the blended model, teachers are required to direct and coach students in how to practice effectively. They monitor, measure and analyze the effectiveness of practice activities by using the following dimensions:

12.5.2.1 Study Days

The amount of practice or total practice time is our first dimension. Without enough practice, progress will be slow. Analyzing from the 1,888 participants studying data, the average study days in one school year was 94 days. The study frequency was 3 days per week, which is our setting standard. From Fig. 12.7, we can see 48 % students study days were more than 90 days, and 10 % students were lower than 70 days (refer to Fig. 12.7 for students study days).

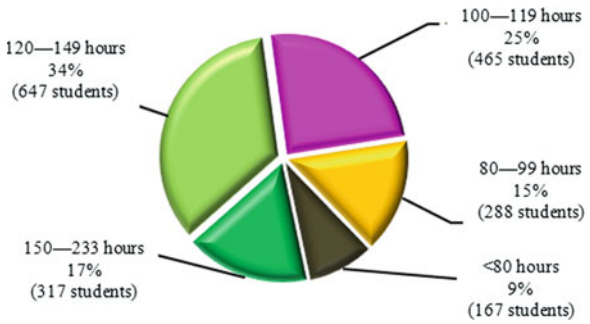


Fig. 12.8 Proportion of students total study hours

12.5.2.2 Study Hours

Another dimension is the frequency of practice. From neuroscience, we know that short, frequent, practice sessions are more effective than long, infrequent sessions. According to the DynEd Records Manager, the average of all participants total study time was 122.5 h. Students study time per week was 4 h, 1.3 h per day. 17 % of the students' total study hours were more than 150 h in one school year, 34 % of the students' study hours were among 120–149 h, 25 % were in 110–119 h and 15 % were in 80–99 h. There are only 9 % of the students total study time were less than 80 h (refer to Fig. 12.8 for students study hours).

12.5.2.3 Study Score (SS)

The third dimension is the quality of practice activities. The quality of a practice activity depends on the action of the students. Recording and comparing native model is an effective way to improve fluency. Study scores indicate how well each student is practicing. But the score it is based on integrated measures including study frequency, study activities, test scores, and comprehension as well as a consideration on students' negative study patterns, such as inappropriate use of text support when developing listening comprehension. It is a measure of how well and how often a student is using a course.

New Dynamic English is the core course of DynEd with a score requirement no less than 4. According to the data collected, the mean study score of NDE was 4.0 with 17 classes got 6.0 or higher and nearly half classes' study score were above the average (See Fig. 12.9).

12.5.2.4 The Quality of the Language Being Practiced

The last dimension of practice is the quality of the language being practiced. Language acquisition takes time and lots of effective practice. Figure 12.10 illustrates the average sentences practiced in different types of exercise in the courseware. All these data were collected from the DynEd Records Manager.

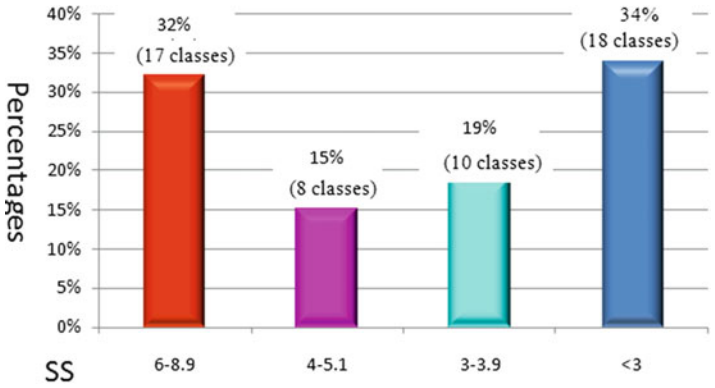


Fig. 12.9 Proportion of average classes study score of NDE

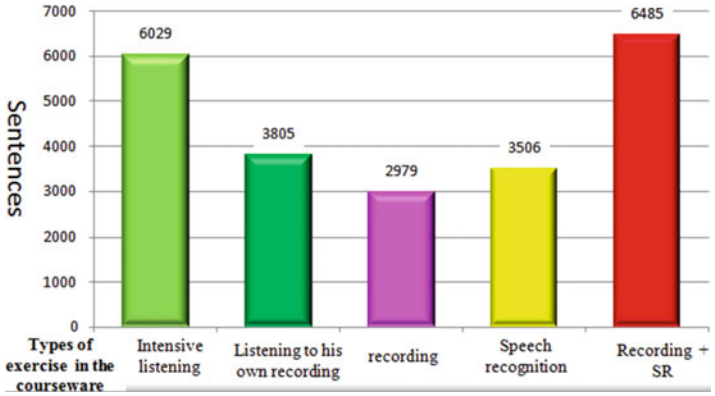


Fig. 12.10 Average number of sentences practiced in different types of exercise

All the students did DynEd courseware learning effectively. On average, each student did 6,029 sentences of intensive listening, 6,485 sentences of recording and speech recognition and 3,805 sentences of listening to his own recording. As we mentioned, after the intensive listening, student should record his sentence, listen to his own recording and compare it with the model. That is an effective way to improve fluency. In the real practice, we find that the students' average number of listening to his own recording is not enough (refer to Fig. 12.10).

12.5.3 Backwash for the Learning and Teaching

We analyze the backwash for the learning and teaching in three ways. One is analyzing the study process of students whose PT level was 0.5 in the 1st PT,

Table 12.6 2nd PT level of 664 students (1st PT = 0.5) and related learning process

2nd PT level	Study days	Study hours	Study score
0.5	93	113	2.9
0.7	95	122	4
1.0	96	125	5.3
1.2–1.7	102	135	7
2.0–2.2	103	143	8.8

Table 12.7 Study process of students in the different growth rate of 2nd PT

2nd PT change	Study days	Total time (h)	Study score
No change	90	115	3.4
<0.5	94	121	4.7
0.5–1.0	96	124	5.4
>1.0	97	137	7.3

another is analyzing the study process of students in the different growth rate of 2nd PT and the last one is the questionnaire.

12.5.3.1 Study Process of Students in the 1st PT Level of 0.5

About 664 students' 1st PT level was 0.5, that was a very representative beginning level. Hence, we did simple observation on their study method and study effect. After one school year's effective learning, 84 % students have made various progress, and 53 % of them have achieved 1.0 in the 2nd PT.

Table 12.6 shows those 664 students in different 2nd PT level and related learning process. Study days and study hours illustrate that the more time students learned, the higher 2nd PT Level got. The study score gives us an important clue, that is, when other variables are controlled, the study score can predict the study effect in some way. Study score means the quality of the language being practiced. Language acquisition takes time and lots of effective practice. The higher study score, the better PT level.

12.5.3.2 The Study Process of Students in the Different Growth Rate of 2nd PT

Besides analysis of the study process of students whose 1st PT level was 0.5, we can analyze the different growth rate between 1st and 2nd PT in the dimension of study method. We divide these growth rate changes into 4 groups, no change, growth rate lower than 0.5, growth rate between 0.5 and 1.0 and growth rate higher than 1.0 respectively.

Table 12.7 shows the detailed learning process of students in the different growth rate of 2nd PT, such as study days, total study time and study score. Students in four groups didn't make much difference on study days and total study time.

Table 12.8 Descriptive statistics of the questionnaire

	N	Mini	Maxi	Mean	Std. deviation
Q1 I feel confident when I communicate in English	189	1	7	4.67	1.502
Q2 My listening ability improves a lot	189	1	7	4.99	1.393
Q3 My pronunciation improves a lot	189	1	7	4.83	1.260
Q4 My oral expression ability improves a lot	189	1	7	4.59	1.433
Q5 My writing ability improves a lot	189	1	7	3.86	1.503
Q6 I can make some communication in English after 1 year's learning	189	1	7	4.50	1.424
Q7 The intelligent tutor of DynEd can give me immediate feedback and monitor my study	189	1	7	5.02	1.663
Q8 Using key buttons in a correct way is very important to effective study	189	1	7	5.21	1.522
Q9 Teacher's analyzing data is very necessary	189	2	7	5.58	1.247
Q10 Classroom teaching based on courseware learning is very important	189	1	7	5.23	1.378
Q11 I am satisfied with the teacher's courseware tutor	189	2	7	5.70	1.206
Q12 I am satisfied with the oral activities in the classroom-teaching	189	2	7	5.71	1.286
Q13 I study hard on DynEd study	189	1	7	5.30	1.292
Q14 I am satisfied with my DynEd study	189	1	7	4.92	1.381
Q15 I hope the lab opening time longer, then I can make more DynEd online learning	189	1	7	5.12	1.508
Q16 I hope I can have more time to study DnyEd	189	1	7	4.51	1.522
Q17 I think DynEd courseware is very novel and interesting, which can motivate me to study hard	189	1	7	4.62	1.488
Q18 I want to continue studying DnyEd to improve my English ability	189	1	7	4.97	1.803
Valid N (listwise)	189				

But the study score influenced a lot of on the growth rate. We can justify that the study score of core course NDE and the growth rate is positively related. If students want to increase the PT level of 0.5 or higher, their study score should reach 5.4 or more. That further validates the accuracy of our initial requirements of the study score should be 4 or more.

12.5.3.3 Questionnaire

At the end of the second semester, we did a questionnaire among 200 students randomly selected from 8 different institutes. Altogether 189 valid questionnaire papers were retrieved. There were 18 items in the questionnaire including students learning effect, teachers teaching effect and the views about DynEd. Response format is 7-point Likert-type scale (strongly agree = 7, strongly disagree = 1). The descriptive statistics by SPSS 17.0 is in table 12.8.

From questionnaires calculations, we can see the mean of each item was more than 4 with the only exception on Q5 about writing ability, which was the lowest. We can draw the conclusion that most students were satisfied with DynEd study, especially with the work of teachers. The mean scores of Q9, Q10, Q11 and Q12 are more than 5 with the mean score of Q11 and Q12 the highest of all the items. This shows students gave much affirmation on teachers' tutoring and classroom-teaching.

But the standard deviation of all the items was more than 1, which illustrates the data was scattered. We cannot exclude the possibility that some students didn't like studying English, especially in this DynEd blended model. DynEd has a super strong data platform – DynEd Records Manager. It can monitor all the learning process of students' courseware learning. Some students said what they have studied, how they have studied and all the learning process are all transparent to their teachers. This transparency gave them too much pressure. On the contrary, some students preferred this blended learning model and they thought highly of DynEd courseware. In their opinion, DynEd placement test gave them a suitable starting point and the Dyned CALL gave them much freedom, for example, they could learn the courseware at their own learning speed and they could listen to or record a difficult sentence as many times as they needed, which was impossible in the traditional classroom. They liked the way DynEd reported their study effectiveness with the "Intelligent Tutor", individually adapted the level of difficulty of their practice activities with the "Shuffler", and evaluated their progress with frequent Mastery Tests. These students enjoyed this blended learning model, and of course, they made much more progress in English proficiency than others.

Though the 4-skills path of language learning within the DynEd system seems quite effective in the improvement in listening, speaking, how to improve students' reading and writing ability needs us further consideration.

12.6 Conclusions

In order to improve students listening and speaking effectively, Ningbo Polytechnic has adopted DynEd courseware to assist students learning English for 3 years. There are nearly 6,000 students have benefited from this blended learning model. This study was a primary attempt to investigate the effectiveness of DynEd CALL by comparing 1st and 2nd Placement Test and Speaking Test as well as the backwash of the tests, using 1,888 non-English major freshmen of 2,009 as the participants. The major contribution to the present college English teaching and assessment is that it is fully proved again that English is a skill, not knowledge and it needs enough practice to make perfect. Besides this teaching concept, a blended model was introduced. It has also confirmed that effective practice leads to language automaticity. DynEd CALL is effective and the award-winning, multimedia content keeps students on task and engaged.

Despite the above-mentioned positive findings, some limitations need to be acknowledged. Firstly, this study was conducted in large-scale and it needs further investigation on some individuals from different PT levels. Secondly, the data needs more professional analyze such as correlation degree, validity and reliability and so on for a more detailed answer on students progress. Thirdly, it also needs further study on testing the hypothesis that oral fluency can benefit writing ability. That will be our further research on this topic.

Acknowledgements The authors are deeply grateful to General Manager Sophie Long and the senior trainer Robert Hsu in DynEd China Ltd. for their valuable help and suggestions.

The research is supported by the Supervisory Committee of English Language Teaching in Vocational Higher Education of Ministry of Education as the key project on English teaching (Project [2011]22).

References

Chen, H. (2011). Research on construction of formative evaluation in English teaching in higher vocational college. *Journal of Nanchang College, 1*, 139–140.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*, 15–37.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language & the brain*. NY: WW Norton.

Gu, X. (2005). Positive or negative? An empirical study of CET washback on college English teaching and learning in China. *ILTA Online Newsletter*. Retrieved from http://www.iltaonline.com/newsletter/02–2005oct/. Accessed 5 Sept 2007.

Gu, G. (2010). A quality analysis of internal college English test items. *Journal of Jiaxing University, 7*, 119–125.

Gu, W., & Liu, J. (2005). Test analysis of college students communicative competence in English. *Asian EFL Journal, 7*, 118–133.

Guo, Q. (2003). Formative assessment and college English teaching and testing. *Tsinghua Journal of Education, 10*, 103–107.

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

Heaton, J. B. (2000). *Writing English language tests*. Beijing: Foreign Language Teaching and Research Press.

Hebb, D. (1949). *The organization of behavior*. New York: John Wiley and Sons.

Hughes, A. (2000). *Testing for language teachers* (pp. 60–62). Beijing: Foreign Language Teaching and Research Press.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Knowles, L. (2004). On the cusp: New developments in language. *Teaching ESL Magazine*, Issue 40, July/August.

Knowles, L. (2008). Mind Blocks. *Language Magazine: Journal of Communication and Education, 8*, 28–34.

Li, J. (2002). The current college English test in China: Problems and thoughts. *Foreign Language Education, 23*(5), 33–38.

Ma, S. C, (2012, March 29). Vocational education system needs top level design. *Guangming Daily*.

Ministry of Education Department of High Education. (2006). *Basic requirement of English teaching for higher vocational education*. Beijing: Higher Education Press.

National College English Testing Committee. (2006). *Syllabus for College English Test*. Shanghai, China: Shanghai Language Education Press.

Pinker, S. (1997). *How the mind works*. Norton & Company: New York. W.W.

Restak, R. (1994). *The modular brain*. New York: Macmillan.

Yang, J. (2008). The study on the current situation of university English final test. *Journal of Liaoning Administration College, 10*, 179–180.

Yuan, J. Y. (2012). Preliminary research on the testing model of the final examination for non-English majors. *Journal of Jilin Agricultural Science and Technology University, 3*, 106–109.

Zhang, B. M. (2011a). Construction of multiple evaluation system in higher vocational college English teaching. *Chinese Vocational and Technical Education, 11*, 35–39.

Zhang, H. X. (2011b). Research on vocational college English test reform. *Journal of Changchun Education Institute, 4*, 134–135.

Zhou, Y. M. (2002). Perspective on university English teaching quality – case analysis. *Foreign Language World, 6*, 71–78.

Chapter 13
Foreign Language Aptitude Components and Different Levels of Foreign Language Proficiency Among Chinese English Majors

Lanrong Li

Abstract This study aims to explore the relationship between language aptitude components and different levels of English proficiency among Chinese English majors. Sixty-four second-year English majors from a university in Beijing participated in the study. An aptitude test composed of three subtests of Pimsleur Language Aptitude Battery and two self-developed subtests was administered to the participants. The students' scores on two national English proficiency tests (Test for English majors – Band 4 and Band 8, abbreviated as TEM-4 and TEM-8) were used as measures of their English proficiency. Correlational analysis, multiple regression analysis and *t*-tests were conducted. Results showed that different aptitude components had different relationship with different levels of language proficiency. Regression analysis showed that two aptitude components (sound discrimination and memory for text) were significant predictors of both of the students' TEM-4 and TEM-8 scores, while only inductive language learning ability could significantly predict their TEM-8 scores. Further analysis showed that students with higher and lower TEM-4 and TEM-8 scores also differed significantly in different language aptitude components. The results lend support to Skehan's hypothesis (A cognitive approach to language learning, Oxford University Press, Oxford, 1998) that different language aptitude components play different roles in second language acquisition when the learner is at different levels of proficiency.

Keywords Foreign language aptitude components • Memory for text • Language analysis • Sound discrimination • Foreign language proficiency

L. Li (✉)
English Department, Beijing Normal University, Beijing, China

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_13,

13.1 Introduction

Foreign Language (FL) aptitude is widely considered as one of the central individual variables in second language (L2) learning (Skehan 1989). Abundant research has found high correlations between foreign language aptitude and various forms of L2 attainment, and the coefficients were usually around 0.4–0.6 (see reviews by J. B. Carroll 1981; Sawyer and Ranta 2001; Dornyei and Skehan 2003). It is believed that foreign language aptitude may account for the difficulties of unsuccessful language learners and the advantages of successful language learners. However, according to J. B. Carroll (1962, 1981, 1990), FL aptitude is not a unitary concept, but is composed of several components, which are phonetic coding ability, grammatical sensitivity, inductive language learning ability, and associative memory.

The componential view of aptitude suggests that learners may draw upon different aptitude components at different developmental stages of L2 proficiency. There is also some evidence (Wesche 1981; Skehan 1986) showing that learners have their strengths and weaknesses in their FL aptitude profiles, and different aptitude components may account for the failure and success of FL learning (e.g. Pimsleur et al. 1963; Ioup et al. 1994; Dekeyser 2000; Rysiewicz 2008). Based on a review of studies on successful and unsuccessful FL learners from the perspective of language aptitude, Skehan (1998) hypothesized that aptitude components play different roles at different proficiency levels and proposed a diagram showing explicitly the relationship between them. His hypotheses make it possible to adopt different instructional methods to facilitate L2 learning at different developmental stages and identify those L2 learners with learning difficulties or stronger potentials.

However, not many studies have attempted to test Skehan's hypotheses. The small number of studies which touched on this issue also generated inconsistent findings (Ma and Wang 2011; Winke 2005; Hummel 2009), especially on the role of grammatical sensitivity and memory components. Due to lack of sufficient empirical evidence, we know even less about the role of inductive language learning ability in L2 development. Thus, the aim of this present study is to test Skehan's hypotheses on the relationship between aptitude components and different levels of L2 proficiency. This article presents a study in which FL proficiency at different developmental stages and FL aptitude components formed the two main variables. In the first section, background about FL aptitude as well as its measurement tools is provided, followed by a literature review on relevant studies on aptitude components and FL proficiency at different levels. The last two sections describe the study and report the major findings.

13.2 Foreign Language Aptitude and Its Measurement Tools

Foreign language aptitude refers to "the individual's initial state of readiness and capacity for learning a foreign language, and probable degree of facility in doing so" given the presence of motivation and opportunity (J. B. Carroll 1981, p. 86). It is

Table 13.1 Structure of the MLAT and the PLAB

Aptitude test	Subtest	Components measured
MLAT	I Number learning	Memory & phonetic coding ability
	II Phonetic script	Phonetic coding ability
	III Spelling clues	Phonetic coding ability & vocabulary
	IV Words in sentences	Grammatical sensitivity
	V Paired associates	Associative memory
PLAB	I Grades in major subjects	
	II Interest	Motivation
	III Vocabulary	Knowledge of vocabulary
	IV Language analysis	Inductive language learning
	V Sound discrimination	Auditory ability
	IV Sound-symbol association	

regarded as a cognitively based learner characteristic that controls the rate of progress the learner will make in foreign language learning. Based on his early research on language aptitude, J. B. Carroll (1981) proposed the following four aptitude components:

1. phonetic coding ability – sound-symbol association ability. An ability to identify distinct sounds, to form associations between those sounds and the symbols representing them, and to retain these associations.
2. grammatical sensitivity – the ability to recognize the grammatical functions of words (or other linguistic entities) in sentence structures;
3. rote learning ability for foreign language materials – the ability to learn associations between sounds and meanings rapidly an deficiently, and to retain these associations; and
4. inductive language learning ability – the ability to infer or induce the rules governing a set of language materials, given samples of language materials that permit such inferences. It is the ability to extract syntactic and morphological patterns from a given corpus of language material and to extrapolate from such pattern to create new sentences (Carroll 1981, p. 105).

The MLAT (Modern Language Aptitude Test), developed by J. B. Carroll and Sapon (1959/2002), was intended to test these aptitude components except inductive language learning ability, which was only weakly and indirectly measured (J. B. Carroll 1981). However, this aptitude component was well represented by the third subtest of PLAB (*Pimsleur Language Aptitude*, Pimsleur et al. 2004). However, PLAB does not have a subtest for testing memory ability, and different from the MLAT, it also takes motivation and average grade points of other subjects as part of aptitude. Besides, the MLAT was designed for adult literate native-speakers of English, while PLAB was designed for native-English speakers aged between 13 and 19. Their structures are presented in Table 13.1

As can be found, there is no one-to-one correspondence between subtests of the MLAT and PLAB and the aptitude components they are intended to measure. Another thing worth noting is that the last two subtests of PLAB measure auditory

ability according to the PLAB manual (Pimsleur et al. 2004). In J. B. Carroll's (1981) model, these two subtests actually tap into phonetic coding ability. However, PLAT-V Sound Discrimination may involve more of the ability to associate sound patterns with their meaning than the ability to distinguish between different sounds (J. B. Carroll 1981; Skehan 1989), as J. B. Carroll (1962) also found that the test of sound discrimination was not predictive of L2 success. In this sense, PLAT-V Sound Discrimination may actually measures two subcomponents: the ability to identify sounds and memory and it may be the latter that gives the validity of this subtest (J. B. Carroll 1981).

Different as the two aptitude tests are, they are complementary rather than contradictory. Since their publication, the two language aptitude tests have been widely used among researchers and have "proved to be strong predictors of language learning success" (Kiss and Nilolov 2005, p. 107). Despite their high predictive validity, some researchers (J. B. Carroll 1990; Robinson 2005) also point out that traditional language aptitude tests may not be able to predict L2 success at the advanced level. Other components like working memory have been proposed as potential aptitude components that predict the success at the advanced level of L2 proficiency (e.g. Miyake and Friedman 1998; Sawyer and Ranta 2001; Robinson 2002, 2005; Dornyei 2005; Erlam 2005). Although the present study did not take working memory as a variable, it is still necessary to provide sufficient evidence to show the limitations of memory ability measured by traditional aptitude test.

13.3 FL Aptitude Components and L2 Learning at Different Levels of Proficiency

By reviewing relevant research on successful and unsuccessful language learners, Skehan (1998) puts forth a diagram showing the relationship between relative importance of the different language aptitude components and different levels of FL language proficiency (as shown in Fig. 13.1).

As can be seen, at the early developmental stage of L2 proficiency, phonemic coding ability is of great importance. However, subsequently, it reaches a plateau, indicating that it only contributes to foreign language learning at the initial stage and once the learner achieved a certain level of foreign language proficiency, it appears not so important. Thus, it is less able to distinguish among learners with relatively advanced foreign language proficiency than the other two components. Language analytic ability has a linear relationship with L2 success, suggesting that it is important at each stage of L2 development. This is basically the same with memory except at exceptionally advanced levels. However, as Skehan (1998, p. 218) admits that "more research is needed, of course, to make this diagram more than a convenient schematic representation", which is also the motivation for the present study. Actually, there is already some partial evidence that lends support to the hypotheses.

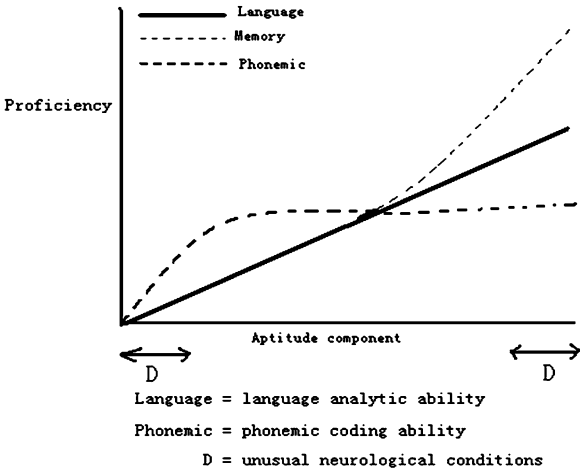


Fig. 13.1 The relationship between aptitude components and proficiency level (Adapted from Skehan 1998, p. 217)

Phonetic coding ability was the component that attracted much attention from Pimsleur et al. (1963) found that auditory ability was important in distinguishing underachieving students from those who showed no difficulty and poor auditory ability often accounted for intelligent or motivated students who seemed unable to learn a foreign language. Sparks and his collaborators (Sparks and Ganschow 1991, 1993, 2001; Sparks et al. 1992; Ganschow et al. 1998) have done a great deal of research on this component. They explored the relationship between phonetic coding ability and measures of mild dyslexia. They noticed that those adolescent or adult novice language learners with low levels in phonological and syntactic coding of their native language tended to experience difficulty in FL learning.

In terms of inductive language learning ability and grammatical sensitivity, Skehan (1998) suggests that the two meld into language analytic ability. Skehan (1986) investigated the relationship between language success and the components of aptitude (language analytic ability and memory ability). By using cluster analysis, he identified patterns in aptitude score profiles associated with success in Arabic and found that some successful language learners relied on their strong language analytic abilities while other successful learners drew more upon their good memory ability. Dekeyser (2000) found that those few adult immigrants who scored within the same range of grammaticality judgment test as child arrivals all got very high scores on MLAT-IV Words in Sentences. Therefore, it can be concluded that language analytic ability may play an important role among adult FL learners and may be related to higher language proficiency.

Memory has received the most interest and attention from researchers and has also undergone considerable change in concept. J. B. Carroll (1981) emphasizes the associative memory between sound and symbols, but Skehan (1982, as cited in Skehan 1998) failed to find any relationship between associative memory and

language learning success. Thus, Skehan (1989, 1998) posits that memory for unfamiliar materials and the ability to impose organizational structures on new materials could better predict the success of language learning. However, Harrington and Sawyer (1992) and Robinson (Robinson 2002) found the relationship between working memory, measured by Deneman and Carpenter's (1980) reading span test, and language learning performance. Skehan (2002) interprets the findings by arguing that in his early research, only a simple span test was used, "which did not require the strong executive working memory, but only rehearsal", while "reading span test is designed to provoke significant computation within working memory" (pp. 75–76). Thus, Skehan concludes that the key issue might be the need to "operate upon the material that is held in immediate memory" (2002, p. 76). Harley and Hart (1997) also failed to find a relationship between associative memory and French proficiency of adolescent English learners but they found that a test of memory for text was correlated significantly with L2 success of early immersion learners (who began a French immersion program from Grade 1).

With regard to the relationship between language aptitude components and different levels of L2 proficiency, there are relatively fewer studies. Winke (2005), who studied foreign language aptitude (measured by MLAT) and working memory of adult English learners of Chinese, found that associative memory, grammatical sensitivity, and phonological memory were associated with learning at the beginning level of Chinese while only phonological memory was associated with learning at the advanced level. Hummel (2009) studied the relationship between aptitude components, working memory and L2 proficiency of adult learners of French at a relatively advanced level and found that grammatical sensitivity and phonological memory could significantly predict the learning of this group. When this group was divided into different groups based on a median split, however, only phonological memory predicted L2 learning of lower-proficiency subgroup, while none of the aforementioned variables predicted L2 learning of higher-proficiency subgroup. Ma and Wang (2011) also studied the relationship between aptitude, working memory and L2 reading among English majors. They found that though all MLAT subtests (except MLAT-V Paired Associates) were correlated with measures of English reading, only MLAT-I Number Learning could predict English reading of the subgroup with lower reading scores while MLAT-IV Words in Sentences and working memory could significantly predict English reading of the subgroup with higher reading scores (the group was also classified by a median split).

In sum, Skehan's hypotheses did not stimulate much empirical research. Existing studies suggest that phonological coding ability seem to play an important role at the initial stage of L2 learning. However, with regard to the role of memory and language analytic ability, relevant studies have not produced consistent findings. Furthermore, studies exploring the relationship between aptitude components and different levels of L2 proficiency all used the MLAT. We know much less about the role of inductive language learning ability and memory and phonetic coding ability measured by subtests of PLAB. The present study aims to

address this issue. Specifically, the study aims to answer the following three questions:

1. What is the relationship between foreign language aptitude and its components and foreign language proficiency among Chinese English majors?
2. Which language aptitude components could predict English majors' proficiency at different stages of development?
3. What are the differences in language aptitude components between higher-proficiency and lower-proficiency groups at different stage of L2 development?

13.4 Methodology

13.4.1 Participants

Sixty-four second-year English majors from a key university in Beijing participated in this study, and 62 finished the aptitude tests. Among them, 9 were males, and 53 were females. Their ages ranged from 18 to 23 with a mean of 20.5 years old. Coming from different parts of the country, 47 students were Hans and the other 15 students were national minorities. On average, they had spent about 7 years learning English when they took the aptitude test. All the participants were from two parallel intact classes enrolled in the same course of Extensive English Reading taught by two different teachers.

13.4.2 Instruments

Owing to practicality, a 45-min aptitude test was designed. Due to lack of appropriate and available tools for measuring Chinese learners' foreign language aptitude, two aptitude subtests were self-developed and the other half adopted the last three subtests of PLAB.

There were five parts in the overall test. Each part was intended to measure different aptitude components. The first two parts were used to measure memory ability, including memory for text and associative memory (see Appendix), which were based on one previous study (Harley and Hart 1997) and modeled on MLAT-VPaired Associates respectively. Inductive language learning ability was measured by PLAB-IV-Language Analysis. Phonetic coding ability was measured by PLAB-V-Sound Discrimination and VI-Sound-Symbol Association. The materials included some slides, a set of booklets, answer sheets, and a questionnaire. The structure of the whole language aptitude test and the points assigned to each part are shown in Table 13.2. There were 143 items in total and each item was worth one point. Detailed information on each subtest of the test is presented in the following section.

The test of memory for text was designed according to Skehan's findings that this aspect of memory ability – the ability "to analyze text, to extract its

Table 13.2 Structure of the language aptitude test

Part	Name	Main focus	N	Task types
1	Memory for text	Memory for text	50	Subjective
2	Paired words	Associative memory	24	Multiple choice
3	Language analysis	Inductive language learning	15	Multiple choice
4	Sound discrimination	Memory	30	Multiple choice
5	Sound-symbol association	Phonetic coding ability	24	Multiple choice
Total	FL aptitude test	/	143	/

propositional content, and remember such content" (Skehan 1989, p. 31)—was significantly correlated with L2 learning success. The test of memory for text in this study was designed on the basis of the work of Harley and Hart (1997).

A narrative Chinese story with about 300 words was selected and prerecorded on the tape. Immediately after listening once to the recording of the story, the students were asked to write down as much of the story as they could recall on the answer sheets within 3 min. The story does not require any special background knowledge to understand. In terms of scoring, the text was deemed to contain 50 information bits based on prepositional analysis. It is generally agreed that a sentence "can be represented by a proposition consisting two or more concepts and some form of relation between them" (D. Carroll 2008, p. 154). For example, sentence "John hit Jack" can be represented as a proposition "hit (John, Jack)". Thus, the following sentences, "Jack was hit by John", "it was John that hit Jack" all can be represented by the same one proposition despite their superficial dissimilarities. The full score for this part was 50 points. A student's score consisted of the number of information bits in each proposition included in their written versions. Another rater was invited to participate in the scoring of this part to decrease the subjectivity in scoring that could have affected the results. The inter-rater reliability coefficient reached .92.

Associative memory test was designed by the author and was modeled on MLAT-4 Paired Associates. First, students were presented a slide showing a list of 24 Icelandic words along with their English equivalents. Then, the students were asked to memorize the English meanings of the 24 words in 2 min, next they needed to choose the corresponding English equivalents for each of them from five choices without looking at the word list in 3 min. Look at the following sample:

fara – go
fara A. road B. ill C. sun D. go E. fly

If the student could choose the choice D by recalling, they would be awarded one point.

With respect to inductive language ability, PLAB-IV Language Analysis was adopted. First, participants read a list of words from a foreign language and the English equivalents of these words. Look at the following sample.

jiban. . . . . .boy, a boy
jojo. . . . .dog, a dog
jiban njojo za. . . . . .A boy likes a dog.

By referring to the above list, they were asked to figure out how the following statement should be expressed in this foreign language.

A dog likes a boy.

This English sentence is followed by four choices and they were asked to choose the correct one.

PLAB-V requires the examinee to differentiate spoken words in an unfamiliar language and learn the meanings of the new words. For the first 15 items, the students were taught 2 words and then must indicate which of two words printed in the test booklet was spoken on the tape. For the subsequent 15 items, the students must choose among all three words and indicate which word was contained in each sentence. PLAB-VI consists of nonsense words based on English syllable structure. The voice on the tape pronounces one of the four words (like trapled, tarpled, tarpdel, trapdel) in each response set, and the students simply indicate which word was spoken.

In terms of the students' English proficiency, their scores on two national English proficiency tests –Test for English Majors Band 4 and Band 8 (abbreviated as TEM-4 and TEM-8) were used as criteria and were collected in 2008 and 2010. TEM-4 contains measures of English listening, vocabulary and grammar, reading and writing while TEM-8 measures English listening, reading, writing and translation. Neither of the two tests contains a subtest measuring speaking. Students are entitled to take the oral test only after they achieve a certain level of written proficiency. TEM-8 is supposed to test students with advanced English levels whereas TEM-4 tests students with intermediate levels. The English Group of the Teaching Guiding Committee for College Foreign Language Majors under the Ministry of Education designed and administered the two tests to second-year and fourth-year English majors respectively in the mainland of China each spring. The two tests have a history of about 20 years and students' scores of the tests serve as important evidence of their English proficiency.

Lastly, a questionnaire was designed to obtain the background information of learners' English learning including sex, age, ethnic background, length of English learning, and their evaluation of the difficulty level of the test.

13.4.3 Procedures

In March 2008, a trial of the test was conducted to test the appropriateness and reliability of the aptitude test. Based on the results of the trial test, the original test was revised. In May, 2008, the revised aptitude test was administered to the participants.

The main study was completed in the last week of May. The whole test took about 45 min and was conducted in one period of normal class. Sixty-four students participated in the study, of whom 62 completed the test. One of the teachers was

invited to invigilate the test with the author to ensure that all the testing procedures be properly followed. Before the test, the purpose and procedures of the test were explained to students in Chinese. Before each of the first three subtests of the aptitude test, the instruction part and a practice session were presented to students on slides. The instruction of the last three aptitude subtests was pre-recorded on the tape.

Students' scores on TEM-4 in 2008 and TEM-8 in 2010 were obtained from the academic office of the department in the school.

13.4.4 Data Analysis

All the data were entered into to one spreadsheet and analyzed by SPSS 16.0. In order to assess the quality of the language aptitude test, descriptive statistics were obtained, including the calculation of means, ranges, and standard deviations of each subscale. Furthermore, for the aptitude subtest of Memory for Text, inter-rater reliability was calculated. For the rest of the aptitude test, item analysis was performed including the difficulty, discrimination index and reliability coefficients by using Cronbach's alpha. To answer the first research question, Pearson product–moment correlational analyses was performed. All tests of correlational significance were two-tailed. The significance level was set at $p < .05$ for this study. To answer the second research question, stepwise multiple regression analysis was conducted. To answer the last question, the participants were first divided into two groups according to the median splits of their TEM-4 and TEM-8 scores. Independent-sample T-test analysis, stepwise multiple regression analyses were employed to compare the differences in language aptitude components between high- and low-proficiency groups.

13.5 Results and Discussion

13.5.1 Students' Performance on English Proficiency Tests and FL Aptitude Tests

Before answering the first question, it is necessary to have a look at the students' performance on language proficiency tests and the aptitude tests. In collecting the data on the participants' English proficiency, one student's TEM-4 score and four students' TEM-8 scores were not available. Table 13.3 presents descriptive statistics about the students' TEM-4 and TEM-8 scores and FL aptitude scores. From the results we can see that the first subtest Memory for Text had a relatively smaller standard deviation, indicating that this subtest was not able to differentiate between the students so well as other subtests.

From Tables 13.3 and 13.4, it can be found that Part 1 Memory for text was the most difficult part of the test. The reason might be that the students were not given

Table 13.3 Descriptive statistics of aptitude test and English proficiency measures

	Min.	Max.	M.	SD
Memory for text	9.50	28.00	17.53	4.55
Paired words	5.00	24.00	18.34	4.27
Language analysis	5.00	15.00	12.35	2.46
Sound discrimination	14.00	30.00	24.16	4.57
Sound-symbol association	2.00	24.00	18.35	3.54
Language aptitude total	63.50	107.50	90.74	10.34
TEM-4	44.00	88.00	69.11	9.71
TEM-8	44.00	82.00	64.23	8.02

Table 13.4 Performance on language aptitude test

	Name	Mean test score (%)	Cronbach's alpha	Mean biserial
Part 1	Memory for text	33.48	/	/
Part 2	Paired words	76.41	.81	0.48
Part 3	Language analysis	82.37	.71	0.44
Part 4	Sound discrimination	80.54	.81	0.39
Part 5	Sound-symbol association	76.48	.75	0.40
Total	Language aptitude	69.86	Part 2–5: .87	Part 2–5: 0.43

enough time to finish this task. The easiest part was Part 3 which measures inductive language learning ability. A possible explanation is that since PLAB is designed for teenagers and the English language involved in this part is rather simple, the inductive language learning ability of the second-year Chinese English majors might have been more developed that the subtest could measure. However, on the whole, we can see that the aptitude test was of appropriate difficulty and reached higher internal consistency and could discriminate students well.

13.5.2 Correlations among Measures of Language Aptitude, Its Components, and L2 Proficiency

Table 13.5 shows the interrelationship between measures of FL aptitude components. First, we can see that although Paired Words and Memory for Text were intended to measure different aspects of memory, their insignificant relationship suggests that they measure distinct cognitive abilities. Paired Words was correlated with Language Analysis significantly at .398 ($p = .001$). This result lends support to J. B. Carroll's (1981) speculation that Language Analysis involves some extent of rote memory ability in that while students were doing the Language Analysis, they also need to establish the association between the newly-learned stimuli and the meanings they represent in a very short time in order to apply the rules underlying the given materials. This result is similar to Harley and Hart (1997)'s findings, in which they found that MLAT-IV Word Pairs and PLAB-IV Language Analysis subtest correlated significantly among both the early and late

Table 13.5 Correlations among measures of language aptitude and its components

	Memory for text	Paired words	Language analysis	Sound discrimination	Sound-symbol association
Paired words	−.077				
Language analysis	.234	.398**			
Sound discrimination	−.155	.032	.267*		
Sound-symbol association	−.123	.111	.200	.444**	
Total	.353**	.526**	.691**	.602**	.577**

Note: *$p < .05$; **$p < .01$

Table 13.6 Correlations between measures of language aptitude and English proficiency

	MfT	PW	LA	SD	SSA	Total	TEM-8
TEM-4	.285*	−.100	.308*	.476**	.339**	.484**	.826**
TEM-8	.299*	.065	.423**	.457**	.251*	.546**	1

Note: *MfT* memory for text, *PW* paired words, *LA* language analysis, *SD* sound discrimination, *SSA* sound-symbol association
*$p < .05$; **$p < .01$

emersion students (at .38 and .54 respectively). Though the participants were different in the two studies, the findings were very similar.

The weak correlation between Sound Discrimination and Language Analysis may be interpreted as that both may involve some extent of rote memory ability. The strong correlation between Sound-Symbol Association and Sound Discrimination was expected as both may have involved the ability to distinguish between and code different sounds.

Table 13.6 shows the relationship between aptitude measures and English proficiency tests. First, the aptitude composite scores were found to be correlated significantly with both TEM-4 and TEM-8 scores. This result is consistent with many of earlier findings (e.g. Gardner and MacIntyre 1992; Ehrman and Oxford 1995; Dai 2006). In terms of the aptitude components, it can be found that all the other aptitude components were correlated with the two English proficiency tests except associative memory measured by Paired Words. This result corroborates some previous findings (e.g. Harley and Hart 1997; Winke 2005), suggesting that associative memory may not play a major role in L2 learning. Even J. B. Carroll (1990) himself admits that he was not confident about the validity of this subtest as its validity seems to vary wildly with different samples.

13.5.3 Differences in Aptitude Components Between Learners at Different Developmental Stages

Table 13.7 shows the results of stepwise regression analysis. As is shown, both Sound Discrimination and Memory for Text were significant predictors of both

Table 13.7 Results of stepwise regression analysis between language aptitude component and TEM-4 and TEM-8

Variable	Step	Predictor	R^2	Adj. R^2	F value[a]	Sig. of F
TEM-4	1	Sound discrimination	.227	.214	17.605	.000
	2	Memory for text	.359	.337	16.492	.000
TEM-8	1	Sound discrimination	.209	.195	15.819	.000
	2	Memory for text	.349	.327	15.786	.000
	3	Language analysis	.394	.363	12.565	.000

Note: Probability for inclusion = .05; probability for exclusion = .01
[a]For equation

TEM-4 and TEM-8. Sound Discrimination, as discussed earlier, may involve more memory ability than the ability to discriminate between different sounds (J. B. Carroll 1981). Sound Discrimination and Memory for Text could jointly explain about 36 and 35 % of the total variance of TEM-4 and TEM-8 respectively. This result suggests the important role of memory ability across different stages of L2 development, thus lending support to Skehan's (1998) hypothesis with respect to memory. However, Language Analysis which was intended to measure inductive language learning ability could only predict the students' TEM-8 scores but not their TEM-4 scores, indicating that this aptitude component becomes more important at the advanced level. Dekeyser (2000) found that those English learners who immigrated to America as adults and achieved native-like proficiency all had outstanding performance on MLAT-IVWords in Sentence. Harley and Hart (1997) also found that inductive language learning ability was associated L2 proficiency of late French immersion learner and speculating that late L2 learners may draw more on their language analytic ability than memory ability. This result seems to give partial support to Skehan's (1998) hypothesis on language analytic ability. One of the interpretations might be that inductive language learning ability might be more associated with productive skills than receptive skills and the former usually develop later than latter ones.

13.5.4 Differences in Aptitude Components Between Subgroups with Higher and Lower TEM-4 and TEM-8 Scores

Tables 13.8 and 13.9 show the differences between the subgroups with higher and lower TEM-4 and TEM-8 scores. First, it can be found that the subgroup with higher TEM-4 scores different from those with lower TEM-4 scores differed significantly at two aptitude component measures – Sound Discrimination and Sound-Symbol Association, suggesting that phonetic coding ability was still important for the learner when they were at the intermediate level of L2 proficiency. However, when the students were in the fourth year of their English study, sound-symbol association was not able to differentiate the two groups any more. Sound discrimination and language analysis became the two factors which could significantly differentiate between the

Table 13.8 Comparison between the two subgroups with higher and lower TEM-4 scores

	Higher TEM-4 (31)		Lower TEM-4 (31)		*t*-test (2-tailed)	
Aptitude	M	SD	M	SD	*t*	*p*
Memory for text	18.06	4.18	17.00	4.90	.920	.361
Paired words	17.90	5.08	18.77	3.29	−.801	.427
Language analysis	12.81	2.48	11.90	2.39	1.461	.149
Sound discrimination	26.00	4.06	22.32	4.35	3.440	.001
Sound-symbol association	19.58	2.17	17.13	4.20	2.886	.006
Total	94.35	9.33	87.13	10.16	2.915	.005

Table 13.9 Comparison between the two subgroups with higher and lower TEM-8 scores

	Higher TEM-8 (31)		Lower TEM-8 (31)		*t*-test (2-tailed)	
Aptitude	M	SD	M	SD	*t*	*p*
Memory for text	18.42	4.38	16.65	4.61	1.553	.126
Paired words	18.45	4.69	18.23	3.88	.207	.837
Language analysis	13.06	2.34	11.65	2.40	2.358	.022
Sound discrimination	26.06	4.06	22.26	4.30	3.586	.001
Sound-symbol association	19.06	3.71	17.65	3.26	1.599	.115
Total	95.06	9.43	86.42	9.48	3.600	.001

two subgroups with higher and lower TEM-8 scores. This result again shows the importance of memory ability across different stages of L2 development. Meanwhile, it also suggests that the ability to associate sounds and their written symbols were not so important once the learner achieves a certain level of L2 proficiency while inductive language learning ability only comes into play at a higher proficiency level.

In order to further find out which aptitude component could significantly predict L2 proficiency of learners at different proficiency levels, another stepwise multiple regression analysis was performed. It can be seen from Table 13.10 that Memory for Text and Sound Discrimination were strongly associated with the subgroup with lower TEM-4 scores, showing the importance of memory at this stage. Perhaps at this stage, the students needed to memorize a larger number of words, phrases, sentences or texts, which might heavily depended upon their memory ability. Previous studies (Skehan 1986; Harley and Hart 1997) suggest that younger learners tend to rely more on their memory ability, while older learners tend to rely more on their language analytic ability. This result seems to suggest that the relationship between aptitude components and L2 learning is not only associated with age but also stages of L2 learning. For the subgroup with higher TEM-4 scores, Sound-Symbol Association was the only significant predictor of their English proficiency. This result is a little bit difficult to interpret and one possible interpretation is that this ability is more related with written than with oral skills and the former might have been developed later than the later.

For the group with lower TEM-8 scores, Language Analysis was the significant predictor of their English proficiency, indicating inductive language learning ability tends to be more important at the lower advanced level. However, for the group

Table 13.10 Stepwise regression analyses: predicting English proficiency at different levels of proficiency from the dimensions of language aptitude

Group	Step	Predictor	R^2	Adj. R^2	F value[a]	Sig. of F value
TEM-4 lower	1	Memory for text	.141	.112	4.776	.037
	2	Sound discrimination	.309	.260	6.267	.006
TEM-4 higher	1	Sound-symbol association	.126	.096	4.192	.050
TEM-8 lower	1	Language analysis	.234	.208	8.880	.006
TEM-8 higher	1	No entry				

Note: Probability for inclusion = .05; probability for exclusion = .01

with higher TEM-8 scores, none of the aptitude components could significantly predict the students' English proficiency. This result supports some previous findings (Winke 2005; Hummel 2009) and could be interpreted as follows. First, it might be due to the limitations of the aptitude test, indicating that some other possible language aptitude components like working memory might be at play at the advanced stage of L2 development. Another possibility is that some other individual variables (e.g. motivation) might be playing a more important role at this stage. Further studies are needed to clarify this issue.

13.6 Conclusion

This study aims to test Skehan's hypotheses regarding the relationship between different aptitude components and L2 proficiency at different developmental stages. On the whole, most of Skehan's (1998) hypotheses were supported. phonetic coding ability was shown to play a major role at earlier stages. However, different from Skehan's hypotheses, memory seemed to be important at all stages except at the rather advanced level, and inductive language learning ability began to play a major role beyond the intermediate level. None of the aptitude components measured by the aptitude test used in this study could predict L2 proficiency at the higher advanced level. Of course, the findings are closely related to the specific aptitude measures used in the study.

Despite these findings, there were some severe limitations. Firstly, the PLAB was designed for native English speakers aged between 13 and 19. Though the level of English language proficiency required for taking the test is assumed to be very low, it is unavoidable that some students with lower levels of English proficiency might have been biased. Therefore, in future studies a language aptitude test suitable for Chinese learners of foreign languages is in urgent need to be developed. Secondly, this study only explored a limited number of aptitude components, it is hoped that in future studies, more aptitude components like grammatical sensitivity and working memory could be included. Thirdly, in terms of language proficiency, this study only took the composite scores. However, it is expected that the results will be more informative if more refined measures of language proficiency like measures of different aspects of language skills and knowledge are taken in future studies.

Appendix: Part 2 Paired Words

1. haf	A. say	B. land	C. sea	D. walk	E. money
2. dv?l	A. sing	B. start	C. bed	D. stop	E. eye
3. lesa	A. land	B. hope	C. read	D. hand	E. write
4. fugl	A. tree	B. bird	C. ant	D. machine	E. light
5. elda	A. fish	B. altar	C. cook	D. cold	E. juice
6. synda	A. speak	B. swim	C. eye	D. heart	E. cake
7. hróp	A. roof	B. fall	C. good	D. call	E. flower
8. andlit	A. fly	B. sing	C. push	D. ground	E. face
9. áta	A. quiet	B. food	C. water	D. clothes	E. shoot
10. hl?ja	A. sky	B. left	C. house	D. laugh	E. shake
11. byssa	A. gun	B. road	C. jump	D. pen	E. night
12. dyr	A. dog	B. lash	C. die	D. animal	E. bed
13. kl?ei	A. cold	B. clothes	C. key	B. night	E. hair
14. skrifbore	A. kite	B. letter	C. hope	D. can	E. desk
15. rétt	A. turn	B. eat	C. people	D. grass	E. pull
16. skera	A. ice	B. break	C. cut	D. ship	E. egg
17. tungl	A. climb	B. moon	C. island	D. jump	E. lamp
18. st?kk	A. march	B. close	C. door	D. jump	E. hear
19. eyja	A. wind	B. strength	C. ship	D. island	E. flow
20. maeur	A. person	B. see	C. cool	D. cry	E. milk
21. vinna	A. travel	B. salt	C. picture	D. job	E. leave
22. útlit	A. top	B. look	C. knife	D. smile	E. drink
23. hlaupa	A. sport	B. day	C. laugh	D. face	E. run
24. vinur	A. map	B. buy	C. friend	D. music	E. cook

References

Carroll, John B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research in education* (pp. 87–136). Pittsburgh, PA: University of Pittsburgh Press.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude.In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Rowley, MA: Newbury House.

Carroll, John B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. Parry & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–19). Englewood Cliffs, NJ: Prentice Hall.

Carroll, D. W. (2008). *Psychology of language* (5th ed.). Beijing: Foreign Language Teaching and Research Press.

Carroll, J. B., & Sapon, S. M. (1959/2002). *Modern language aptitude test*. Bethesda: Second Language Testing.

Dai, Y. (2006). The effects of language aptitude on second language acquisition. *Foreign Language Teaching and Research, 6*, 451–459.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 493–533.

Dornyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah: Lawrence Erlbaum.

Dornyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Oxford: Blackwell.

Ehrman, M., & Oxford, R. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal, 79*, 67–89.

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9*, 147–172.

Ganschow, L., Spark, R., & Jovorsky, J. (1998). Foreign language learning problems: An historical perspective. *J Learn Disabil, 31*, 248–258.

Gardner, R., & MacIntyre, P. (1992). A student's contribution to second language learning. Part 1: Cognitive variables. *Language Teaching, 25*, 211–220.

Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition, 19*, 379–400.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*, 25–38.

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics, 30*, 225–249.

Ioup, G., Boustagui, E., Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition, 16*, 73–98.

Kiss, C., & Nilolov, M. (2005). Developing, piloting, and validation an instrument to measure young learners' aptitude. *Language Learning, 55*, 99–150.

Ma, Z., & Wang, T. (2011). The effects of language aptitude and working memory on L2 reading comprehension. *Shandong Foreign Language Teaching, 32*, 41–47.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy,& L. E. Bourne Jr (Eds.), *Foreign language learning:Psycholinguistic studies on training and retention* (pp. 339–364). Mahawah, NJ: Lawrence Erlbaum.

Pimsleur, P., D. M. Sunderland, and R. McIntyre. 1963. *Underachievement in foreign language learning* (ERIC Document Reproduction Service No. ED018160). Retrieved from ERIC database.

Pimsleur, P., Reed, D. J., & Stansfield, C. W. (2004). *Pimsleur language aptitude battery: Manual*. Bethesda: Second Language Testing, Inc.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfeld and Hernstadt, 1991. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211–266). Amsterdam: John Benjamins.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25*, 46–73.

Rysiewicz, J. (2008). Cognitive profiles of (un)successful FL Learners: A cluster analytical study. *The Modern Language Journal, 92*, 87–99.

Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319–353). Cambridge: Cambridge University Press.

Skehan, P. (1986). Cluster analysis and the identification of learner types. In V. Cook (Ed.), *Experimental approaches to second language learning* (pp. 81–94). Oxford: Pergamon.

Skehan, P. (1989). *Individual differences in second language acquisition*. London: Edward Arnold.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2002). Theorizing and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–93). Amsterdam/Philadelphia: John Benjamins.

Sparks, R., & Ganschow, L. (1991). Foreign language learning differences: Affective or native language aptitude differences? *The Modern Language Journal, 75*, 3–16.

Sparks, R., & Ganschow, L. (1993). Searching for the cognitive locus of foreign language learning difficulties: Linking first and second language learning. *The Modern Language Journal, 77*, 289–302.

Sparks, R., & Ganschow, L. (2001). Aptitude for learning a foreign language. *Annual Review of Applied Linguistics, 21*, 90–111.

Sparks, R., Ganschow, L., Javorsky, J., Pohlman, J., & Patton, J. (1992). Test Comparisons among students identified as high-risk, low-risk, and learning disabled in high school foreign language courses. *The Modern Language Journal, 76*, 142–159.

Wesche, M. B. (1981). Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 119–154). Rowley, MA: Newbury House.

Winke, P. 2005. *Individual differences in adult Chinese second language acquisition: the relationships among aptitude, memory, and strategies for learning*. Unpublished doctoral dissertation, Georgetown University.

Chapter 14
Motivation and Arabic Learning Achievement: A Comparative Study Between Two Types of Islamic Schools in Gansu, China

Juping Qiao, Kassim Noor Lide Abu, and Badrasawi Kamal

14.1 Introduction

In China, Arabic is one of the earliest foreign languages learned by people. During the Han Dynasty (206 B.C.–220 A.D.), Arabic language was spoken and learned by few Chinese people when the relation between China and Arab was established. In the Tang Dynasty (618–907) and the Song Dynasty (960–1279), a lot of Arabs came to China to trade to enhance the bilateral relation between the two countries. During the Yuan Dynasty (1271–1368), Arabic language was widely learned and studied. With the spread of Arab-Islamic culture in different regions of China, Muslim population rapidly increased. From the late Ming Dynasty (1368–1644) to the early Qing Dynasty (1644–1912), Islam reached the maturity stage that brought a large number of Muslim scholars who greatly contributed to the development of Chinese Islamic education. Arabic language emerged as a significant foreign language. This foreign language was studied until today (Li and Feng 1998).

Although Arabic language has been learned informally and formally from the Tang Dynasty till today its comparative importance to other foreign or second languages is almost Negligible. Until now, there are no students learning Arabic at schools, despite the fact that many Muslim children go to some of these schools. In addition, many Chinese Muslim students are still learning Arabic language informally; of approximately 20 million Muslims, only 25000 Muslim students are studying Arabic language as a specialization in public universities and colleges, Islamic universities and colleges, and Islamic Schools either in China or outside China.

J. Qiao (✉) • B. Kamal
Institute of Education, International Islamic University, Selayang, Selangor, Malaysia
e-mail: qiaojuping@gmail.com

K. Noor Lide Abu
Kulliyyah of Dentistry, International Islamic University, Selayang, Selangor, Malaysia

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_14,

14.1.1 Self-Determination Theory and Second Language Learning (LL2)

In the academic and non-academic setting, motivation is a driving force that causes individuals to pursue goals or achieve success. Deci and Ryan (1985) classified motivation as intrinsic and extrinsic motivation based on the degree of self-motivation for learning a second language. They highlighted that language learners are either interested in learning a new language for their own sake (intrinsic) or for rewards (extrinsic).

> Extrinsic motivation refers to motivation that exists because of the presence of "an externally mediated activity or constraint" (Deci 1980, p. 30–31). Extrinsically motivated learners to acquire a second language are to improve their chances of getting a good job. Intrinsic motivation refers to "doing something because it is inherently interesting or enjoyable" (Deci and Ryan 2000, p.61).

They also distinguished between self-determined (intrinsically motivated) and non-self-determined (extrinsically motivated) behaviors through the Self-Determination Theory. This theory offers "a very interesting look at motivation by setting a different agenda for language teachers. Rather than focusing on how people (e.g., teachers in the classroom) can motivate others, the focus should be on how people can create the conditions within which others can motivate themselves" (Keblawi 2009, p.32). More specifically, it is used to explicate L2 motivational issues (Dornyei 2003; Noels 2001a, 2001b; Noels et al. 1999, 2000, 2001). Self-Determination Theory evolved from studies comparing the intrinsic and extrinsic motives that develop to intrinsic and extrinsic motivation (Deci and Ryan 1985), and examines three motivational subscales that are intrinsic motivation (IM), extrinsic motivation (EM), and amotivation (AM). Deci and Ryan (2000) graphically depicted these three motivational subscales and pointed out that motivation to acquire a language evolved from intrinsic and extrinsic motives (Fig. 14.1).

According to this model, reasons for learning a second language can be classified according to the degree to which learners freely choose to learn another language, and the degree to which they are self determined (Deci and Ryan 1985; Noels 2001a, b; Noels et al. 2000, 2001; Vallerand et al. 1997).

Intrinsic motivation is related to second language learning and achievement. More specifically, learners' motivations are self-determined; they become more involved in learning a second language and make efforts to reach the learning goal (Deci and Ryan 1985, 2000). Some previous studies further illustrated the importance of intrinsic motivation on a second language. For example, Noels et al. (2000) revealed that high intrinsic motivation could drive students, registered in English psychology class at French-English bilingual university, to reach a learning goal. Lucas et al. (2010) indicated that "the stronger feelings of intrinsic motivation are related to positive language learning outcomes, including greater motivational intensity, greater self-evaluations of competence, and a reduction in anxiety" (p.19). Similarly, Bakar et al. (2010) found that the effect of intrinsic motivation on Arabic as the second language is significant and positive. Other studies also

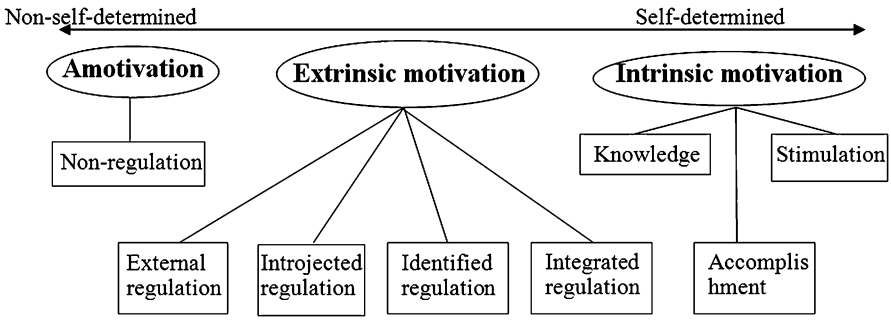


Fig. 14.1 Orientation subscales along the self-determination continuum (Source: Deci and Ryan (2000, p.61))

found a positive relationship between intrinsic motivation and second language (Deci and Ryan 1985, 2008; Dornyei 1994, 1996, 1998a, b; Ryan and Stiller 1991).

Extrinsic motivation could be called a motivation that exists because of external factors causing individuals to acquire a second language (Deci and Ryan 1985, 2000; Noels et al. 2001; Ryan and Stiller 1991). This basic motivation has traditionally been seen as potentially undermining intrinsic motivation. Harter (1981) found that students would lose their natural intrinsic interest in an activity (such as LL2) if they have to do it to meet some extrinsic requirement. Noels (2001a) argued that students learn a second language more actively when teachers give praise and encouragement to them. Bakar et al. (2010) articulated that high extrinsic motivation could motivate learners to acquire a new language effectively. In this regard, LL2 is performed, not for the enjoyment of the activity, but in order to gain a reward or avoid the punishment.

Another motivational subscale is amotivation. It is independent from intrinsic and extrinsic motivation, and it represents the absence of either intrinsic or extrinsic reasons in LL2. Studies have indicated a correlation between amotivation and second language learning (Bakar et al. 2010; Chambers 1993; Dornyei 1998b; Noels et al. 2001; Oxford 1998; Ushioda 1998). They found that amotivation negatively influenced second language learning achievement.

To sum up, three aspects of motivation (intrinsic motivation, extrinsic motivation and amotivation) are closely related to a second language learning and achievement. According to the Self-Determination Theory, "motivation represents one of the most appealing, yet complex, variables used to explain individual differences in language learning" (MacIntyre et al. 2001, p. 462).

14.1.2 Research Purpose

The primary purpose of this study was to explore factors that influence Arabic learning achievement and examine the relative contribution of the different aspects of motivation (Intrinsic Motivation [IM], Extrinsic Motivation [EM], Amotivation

[AM], and Religious Motivation [RM]) on Arabic learning achievement. The secondary purpose of this study was to investigate the relationship between different aspects of motivation (IM, EM, AM, and RM) and compare the level of motivation (IM, EM, AM, and RM) between students from two types of Gansu Islamic Schools in China.

14.2 Research Methodology

14.2.1 Participants

The population of this study refers to 585 male high-school graduates from three Gansu Islamic Schools, who attended Arabic language classes for one or more semesters. 60 % ($n = 348$ [Sampling $= 585 \times 60\ \% = 348$]) of high-school graduate students from the three Gansu Islamic Schools were randomly selected as the sample. From the Lanzhou Islamic College, 177 students were randomly selected from three sections; 75 students from two sections of the Guanghe Arabic School; and from the Linxia Arabic School, 96 students from three sections were selected.

14.2.2 Instrumentation

Finding a good instrument that measures the independent, dependent, and control variables is very important. In this study, the research instrument is employed to answer four research questions. In order to answer the first research question, in-depth interviews with some selected students, a teacher and a rector from two Gansu Islamic Schools were conducted. The interview questions were developed from the subscales of the Self-determination construct proposed by Bakar et al. (2010). As this study examined how religious motivation enhances self-determined behaviors in Arabic language learners, it introduced religious motivation as part of the four motivational subscales within the Self-Determination construct of language learning based on Deci and Ryan (2000), Noels et al. (2000), and Bakar et al.'s work (2010) (Table 14.1).

The items for the seven subscales (Intrinsic Motivation-Knowledge, Intrinsic Motivation-Accomplishment, Intrinsic Motivation-Stimulation, Extrinsic Motivation-External Regulation, Extrinsic Motivation-Introjected Regulation, Extrinsic Motivation-Identified Regulation, and Amotivation) were adopted from Noels et al. (2000). They were essentially statements representing the dimensions underlying French learner motivation in learning English psychology.Bakar et al. (2010) extended the self-determination theory further. They examined six motivational subscales (Intrinsic Motivation-Knowledge, Intrinsic Motivation-Accomplishment, Extrinsic Motivation-Identified Regulation, Extrinsic Motivation-External

Table 14.1 The subscales within the self-determination construct of language learning

Motivation: The driving force for individuals in performing tasks and achieving goals, it can be said to be intrinsic and extrinsic and generally used for humans but also can be used to the causes for animal behaviors		
Religious motivation: It is an important motivational system that closely interacts with the social culture, learners with various degrees of quality that would be relate to Arabic language learning (e.g., I learn Arabic language because it is good for the understanding of Quran)		Highest self –determined
Intrinsic motivation: The source from internal or inherent motivates to face challenges (e.g., performing tasks, engaging activities and achieving goals) for personal satisfaction and enjoyment; IM is most often associated with cognitive and social development. So IM caused behaviors represent high self-determined		
IM-knowledge: Internal constructs such as curiosity, feeling and self-interest, which motive individuals to explore new ideas and develop knowledge for personal satisfaction and enjoyment	I like to learn Arabic language interested by me for the self-satisfaction and enjoyment	
IM-accomplishment: Learners interact with environment in order to feel satisfy for creating the accomplishment and enjoying the process of achieving goals	I am satisfying in acquiring Arabic language and passing the exam	
IM-stimulation: Learners engage in activities or performing tasks in order to experience stimulating sensations such as fun and excitement	I asked questions in Arabic class for the sensation of pleasure stimulated by teacher's praise	
Extrinsic motivation: Learners acquire, learn or create something in order to earn a reward or avoid punishments; EM caused behaviors represent relatively lower self-determined		
Identified regulation: Learner's behaviors are regulated because of personal reasons such as a valuable personal goal. External and somewhat self-determined	I learn Arabic language because it is important for my personal development	
Introjected regulation: Learner's behaviors are regulated due to the external pressure incorporated into them. External and somewhat self determined	I learn Arabic language before exam because good students learn	
External regulation: Learner's behaviors are regulated through external demands or possible rewards. External and least self-determined	I study Arabic language for a good salary later	
Amotivation: Learners with a lack of IM and EM that leads to low motivation in engaging in the activity of Arabic language learning (e.g., I don't have any chance to practice Arabic language without class)		Lowest self-determined

Based on three studies: Noels et al. (2000); Bakar et al. (2010); and Deci and Ryan (2000)

Regulation, Amotivation, and Religious Motivation) on Malaysian learner motivation in learning Arabic language.

Intrinsic Motivation (IM), Extrinsic Motivation (EM), Amotivation, and Religious Motivation (RM) are four major constructs of this study. The four constructs were measured using 45 items in a questionnaire. The items were adapted from previous research (Table 14.2). The instruments, however, were administered in Chinese rather than English or Arabic languages. This is due to two reasons: (1) Chinese language is the learners' mother tongue and would contribute to a better understanding of the items in the questionnaire; and (2) some of the learners were beginners in English and Arabic languages that they would face difficulty in understanding the questionnaire if it were in the English and Arabic languages.

The scale consists of four main sections. The first section gathers factors related to religion (Islam). The part covering three items measures the religious motivation of Muslim students in learning Arabic language. The second section consists of questions that focus on some issues related to Arabic learning outcomes. This part comprises 12 items which are used to measure Intrinsic Motivation (IM) on Arabic language learning achievement. There are three constructs that make up this section: IM-Knowledge, IM-Accomplishment, and IM-Stimulation. IM constructs has been shown to positively relate to Arabic language learning (Bakar et al. 2010). The third part is Extrinsic Motivation (EM) consisting of EM-External Regulation, EM-Introjected Regulation, and EM-Identified Regulation. There are 12 items used to measure EM on Arabic language learning achievement in this survey. The fourth part is Amotivation consisting of 18 items, which measured negative factors on students' motivation in Arabic language learning. The four scales have been proposed to the domain principle in second language learning (Bakar et al. 2010; Deci and Ryan 1985; Noels et al. 2001; Vallerand et al. 1997; Vallerand et al. 1989, 1992, 1993). These researchers used the instrument to measure different aspects of motivation in the second languages. However, some items were found to be irrelevant for this research setting and therefore not included in this instrument. The researcher, thus, develop more suitable items for the research setting.

The 45 items were rated on a 7-point Likert scale with 1 indicating 'Strongly Disagree' and 7 representing 'Strongly Agree'. The possible scores ranged from 45 to 315. Students with a high score indicate a high level of motivation in Arabic language learning while those with a low score have low motivation in Arabic language learning. The items were piloted and 36 items were selected for the actual data collection. The items that were removed were 9 items on amotivation, which were found to be unsuitable.

14.3 Results of the Pilot Study

In order to develop a suitable questionnaire instrument for the research setting, the researcher conducted a pilot study involving 45 senior-school graduates from the Guanghe Arabic School. For the pilot study, high scores (6 and 7) indicate a

Table 14.2 Constructs and sources of questionnaire used

Construct	No of items	Source		
Religious motivation	3 (1, 2, 3)	**Item 1, 2 and 3 are adapted from Bakar et al. (2010)** I want to understand the content of Quran and Hadith better I compel myself to learn the language of Quran I compel myself to understand prayers which are recited in Arabic		
Intrinsic motivation (IM)	12	**Intrinsic motivation** covers three subscales: IM-knowledge, IM-accomplishment and IM-stimulation (Vallerand et al. 1989) Among the 12 IM items, 9 items are adopted from two studies conducted by Noels et al. (2001) and Bakar et al. (2010). Another 3 items are developed by the researcher according to the definitions of IM-accomplishment and IM-stimulation (Vallerand et al. 1989; Noels et al. 2001)		
IM-knowledge	4 (4, 5, 6, 7)	**Items 4, 5, 6 and 7 are adopted from Noels et al. (2001)** For the satisfied feeling I get in finding out new things Because I enjoy the feeling of acquiring knowledge about the second language community and their way of life	**Items 4, 5, 6 and 7 are adopted from Bakar et al. (2010)** I feel satisfy in finding out new things I like to know the differences between Arabic and my own language I am serious about Arabic culture I like to compare and contrast Arab culture and my own culture	
IM-accomplishment	5 (8, 9, 10, 11, 12)	**Items 10 and 12 are adopted from Noels et al. (2001)** For the satisfaction when I am in the process of accomplishing difficult exercises in the second language	**Items 8, 10 and 12 are adopted from Bakar et al. (2010)** For the satisfaction that I feel when I can acquire a second language	**Item 9 is adopted based on its definition** *IM-accomplishment* refers to "the sensations related to attempting to master a task or achieve a goal" (Noels et al. 2001)

(continued)

Table 14.2 (continued)

Construct	No of items	Source		
		For the enjoyment I experience when I grasp a difficult construct in the second language	For the satisfaction when I feel I can accomplish difficult exercises in Arabic For the pleasure that I feel I can grasp the complexity of Arabic grammar	e.g. For the pleasure when I feel I can get good marks in L2 learning
IM-stimulation	3 (13, 14, 15)	**Item 14 is adopted from Noels et al. (2001)**	**Items 13 and 14 are adopted from Bakar et al.'s work (2010)**	**Item 15 is adopted based on its definition**
		For the high feeling that I experience while speaking the second language	Because the lesson are interesting For the high feeling that I feel I can speak in Arabic	*IM-stimulation* refers to motivation based on sensation stimulated (e.g. fun and satisfaction) by performing tasks (Noels et al. 2001)
Extrinsic motivation (EM)	12	**Extrinsic motivation** covers three subscales: EM-identified regulation, EM-introjected regulation and EM-external regulation (Vallerand et al. 1989)		
		Among the 12 EM items, 3 items are developed based on studies conducted by Noels et al. (2001) and Bakar et al. (2010). In addition, 1 EM item was adopted from Noels et al. (2001) and 3 EM items were adopted from Bakar et al. (2010). Another 5 items are developed according to the definitions of three EM subscales (Vallerand et al. 1989; Noels et al. 2001)		
EM-identified regulation	5 (16, 17, 18, 19, 20)	**Items16 and 18 are adopted from Noels et al. (2001)**	**Items 17, 18 and 19 are adopted from Bakar et al. (2010)**	**Item 20 is developed according to its definition**
		Because I think it is good for my personal development Because I choose to be the kind of person who can speak more than one languages	Arabic is good for my personal development Choose to speak Arabic Choose to speak many languages Choose to speak language of the Prophet	*EM-identified regulation* refers to behaviors regulated by external factors, such as salary and job (Vallerand et al. 1989; Noels et al. 2001)

EM-introjected regulation	3 (21, 22, 23)	**Item 21, 22, 23 are developed according to its definition** EM-introjected Regulation refers to individual behaviors which are regulated because of the interaction between internal and external factors. For example, an individual helps others because of his/her personal ethical standards (Deci and Ryan 1985; Vallerand et al. 1989; Noels et al. 2001)		
EM-external regulation	4 (24, 25, 26, 27)	**Items 24, 26 adopted from Noels et al. (2001)**	**Items 24, 26, 27 adopted from Bakar et al. (2010)**	**Item 25 is developed based on its definition**
		In order to have a better salary later on	In order to have a better salary latter on	*EM-external regulation* refers to behaviours is regulated through external demands or possible rewards (Vallerand et al. 1989; Noels et al. 2001)
		In order to get a more prestigious job	In order to get a more prestigious job Because it is a university requirement	e.g., Learning English (L2) is for studying abroad
Amotivation	18 (28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45)	**Items 28, 40 adopted from Noels et al. (2001)**	**Item28, 31, 40 developed from Bakar et al. (2010)**	**Other 15 items are developed based on its definition**
		I cannot come to see why I study a second language, and frankly, I don't give a damn	I don't care if I pass Arabic course	With intrinsic and extrinsic motivation, amotivation comes from internal and external sources that lead individuals to lose interest in performing tasks or achieving a goal (Kasser and Ryan 1993, 1996; McHoskey 1999; Vallerand et al. 1989; Noels et al. 2001)
		I don't know; I cannot come to understand what I am studying a second language	I cannot understand why I am studying Arabic Arabic has no practical importance for my future	

Source: Questionnaire is developed by Noels et al. (2000) (Appendix D), Questionnaire is applied in Bakar et al.'s work (2010)

high level of motivation in Arabic learning, while a low scores (1 and 2) show low level of motivation in Arabic learning. With the use of Principal Component Analysis (PCA) in SPSS irrelevant items were removed from the final questionnaire. After data analysis, it is found that 9 items were found to improve the reliability if deleted SD <1. Thus, 36 items were identified for the research (Table 14.3).

14.3.1 Data Collection and Analysis

Data collection is generally obtaining useful information. The purpose of data collection is to improve the process, not the product it produces. In this study, data collection has two steps: interviews and student surveys.

There are three methods for data analysis utilized to answer the research questions in this study. They are Principal Component Analysis (PCA), the Rasch Model, Pearson's Correlation, Multiple Regression, and *T*-Test. PCA is to remove inapplicable items from the final questionnaire that match for the research setting; the Rasch Model is applied to transform the ordinal data gathered from the questionnaire into interval measures; the Pearson's correlation is employed to examine the relationship between four aspects of motivation (IM, EM, AM, and RM); multiple regression is applied to explore the relative contribution of between four aspects of motivation (IM, EM, amotivation and religious motivation) on Arabic language learning achievement; and *T*-Test is utilized to indicate the difference in four motivational variables (IM, EM, AM, and RM) between students from the two Islamic Schools.

14.3.2 Research Results

RQ1 What factors are perceived to contribute to Arabic learning achievement in the Gansu Arabic Schools and Gansu Islamic College in China?

This research question is mainly concerned with the positive and negative factors that affect students' achievement in the Arabic language. To answer this question, interviews with a rector of an Arabic-school, a teacher of the Arabic language, a Chinese Arabic-school graduate student, and four school students learning Arabic language, were conducted to accrue in-depth information regarding: (a) the positive and negative factors that affect students Arabic achievement; (b) the relationship between religion and Arabic language learning; and (c) suggestions on how to improve students' achievement in the Arabic language.

Table 14.3 Results of the pilot study

Items	N	Mean	SD	Comm.
Religious motivation				
I want to understand the content of Quran and Hadith better	45	4.64	1.82	.764
I compel myself to learn the language of Quran	45	4.64	4.51	.899
I compel myself to understand prayers which are recited in Arabic language	45	2.69	1.70	.860
IM-knowledge				
I like to learn the content of Arabic textbook	45	3.09	1.53	.848
I like to know the differences between Arabic language and my own language	45	2.42	1.39	.750
I am curious about some Arabic culture	45	2.98	1.57	.777
I like to compare and contrast Arab culture and my own culture	45	1.80	1.06	.833
IM-accomplishment				
For the satisfaction that I feel when I can acquire a second language	45	1.91	1.16	.886
For the pleasure that I feel when I can get a good GPA	45	1.96	1.22	.797
For the satisfaction that I feel when I can accomplish difficult exercises in Arabic	45	1.91	1.12	.804
For the pleasure that I feel when I can answer questions in the Arabic class	45	3.16	1.50	.799
For the enjoyment that I feel when I can grasp the complexity of Arabic grammar and morphology	45	1.93	1.14	.727
IM-stimulation				
I feel Arabic lessons are very interesting	45	3.11	1.54	.698
For the high feeling that I feel when I can speak in Arabic	45	1.82	1.01	.745
For the pleasure that I feel when I can learn Arabic	45	1.69	1.33	.822
EM-identified regulation				
Arabic is good for my personal development	45	1.71	1.22	.842
I choose to speak Arabic	45	4.60	4.52	.811
I choose to speak many languages	45	3.02	1.53	.832
I choose to speak the language of the Prophet	45	2.56	1.42	.774
I choose to speak the language of the Quran	45	3.18	1.67	.833
EM-introjected regulation				
I work hard to learn Arabic for my parents' pleasure	45	2.36	1.38	.839
I work hard to learn Arabic for my teachers' pleasure	45	2.22	1.18	.876
I feel guilty if I don't learn Arabic when good students are supposed to do	45	2.93	1.54	.875
EM-external regulation				
In order to get a high-salary job later on	45	2.44	1.53	.720
In order to study abroad later on	45	2.91	1.88	.862
In order to get a more prestigious job later on	45	5.33	1.49	.790
Because it is a school requirement	45	5.37	1.43	.751
Amotivation				
Arabic has no practical importance for my future	45	1.62	.886	.900
It is hard to find Arabic teacher after lessons	45	4.87	1.87	.840
I don't understand what teacher is speaking during the class	45	1.58	.811	.979
I don't care if I cannot pass my Arabic course	45	1.62	.936	.717
I just don't understand how to learn Arabic effectively	45	5.02	1.71	.707

(continued)

Table 14.3 (continued)

Items	N	Mean	SD	Comm.
I cannot accomplish Arabic assignments on time	45	1.62	.860	.886
Arabic teacher is not interesting	45	1.57	.813	.971
I never answered questions in the lessons	45	5.31	1.74	.752
I never asked questions in the lessons	45	3.11	2.05	.815
Arabic class is boring	45	1.60	.809	.953
I feel nervous in Arabic class	45	5.84	1.56	.714
I don't have chance to practice Arabic without classroom	45	5.40	1.75	.766
I cannot understand why I am studying Arabic	45	1.60	.837	.964
My classmates always change	45	3.87	1.89	.768
We always change Arabic teacher	45	1.49	.789	.872
My Arabic grammar is not good	45	3.13	2.13	.699
My Arabic morphology is not good	45	3.22	1.99	.713
Arabic class is so difficult for me	45	1.53	.786	.961

Descriptive statistics of 45 motivational items (N, Mean, SD, Communalities)
Community (p) $>$ .60, SD $>$1

14.3.3 Factors Affecting Students' Achievement in Arabic Language

The analyses of the respondents' interviews show that all the respondents have put great emphasis on several external and internal factors that help students achieve better in the Arabic language. This finding is supported by the respondents' answers to the interview question, "*What factors do you think would positively motivate students to learn Arabic language and achieve good learning outcomes*?" The intrinsic positive factors cited include: high motivation or interest in learning the content of Arabic culture; positive attitude towards learning the Arabic language; high satisfaction in learning the Arabic language or in acquiring a new language; enjoying learning the Arabic language; and feeling that Arabic language is interesting. For example, the Chinese Arabic school graduate mentioned that among the internal reasons of learning Arabic are the high interest in and the positive attitude towards learning Arabic language. He stated that students' achievement in the Arabic language.

> According to my learning and teaching experience, sufficient effort, high interest, and positive learning attitude are greatly and intrinsically contributed to students' good learning outcomes. For sure, a good job and future study outside Chinese Universities arouse students' high motivation to learn Arabic language.

The respondents also indicated that there are some external factors that make students learn the Arabic language better. These include studying abroad; finding a good or more prestigious job; personal development; appreciation given to Arabic learners and speakers; and family's and teachers' encouragement. This is supported by the respondents' awareness of the instrumental use of the Arabic language such as getting a good job, studying abroad, understanding the Arabic culture, and communicating with the Arabic speakers, as shown in the following statements by the Arabic teacher

> Ok. Many external factors are related to students' motivation in Arabic language achievement. For example, students prefer to choose speaking the Arabic language...The language of the Prophet and Quran could contribute to high Arabic learning achievement. Students would try their best to learn the Arabic language so that they could get a high-salary or a prestigious job later on. Some learn Arabic language for studying abroad in future.

On the other hand, the analyses of interviews showed that there are external and internal factors that negatively influence students' achievement in the Arabic language. The internal factors include lack of motivation to learn Arabic or loss of interest in learning Arabic; language anxiety; nervousness; and boredom. This is supported by the respondents' answers when they were asked about the factors that negatively affect students' achievement in Arabic language. For instance, the Arabic school rector said that lack of motivation leads to low marks in Arabic.

> Indeed, the purpose of studying here is not only to learn, but also to improve one's personality. So, these students lack motivation to learn Arabic language that leads to low GPA.

One Arabic learner added that feeling anxious, nervous or bored in the Arabic classes affect students learning the language.

> The teachers of Arabic language as well as the Arabic classes are very boring for students. They do not ask or answer questions in Arabic classes. Even few students do not know why they are learning the Arabic language. These factors might lead us to lose interest to learn Arabic language. I think I do not have the language gift although my GPA is ok. I always feel nervous in the Arabic class and reject to ask or answer questions.

The external factors that negatively affect students' achievement in the Arabic language include students' poor educational qualification; Arabic classes being too difficult and uninteresting; inappropriate teaching methodology; lack of participation or engagement in academic activities; limited uses of Arabic after graduation; lack of understanding of the importance of Arabic language; and the constant change of teachers of Arabic. Respondents also explicated that some students are not qualified enough to enter religious or Arabic schools. They are sent to these schools either because other schools have rejected them or their parents do not want their children to be in the streets and be involved in illegal activities. The rector said that

> Why students come, directly lead to academic failure in the Arabic language. Many students choose to come here when they are rejected to join other schools, universities or colleges. Parents are scared of their sons being involved in illegal activities in the society, and so they send them to the religious schools. Indeed, the purpose of studying here is not only to learn, but also to improve one's personality.

Moreover, students do not understand Arabic classes because teachers do not apply appropriate methodologies, which help students understand and be more interested in learning Arabic.

14.3.4 Relationship between Religion and Student Achievement in Arabic Language

In order to elicit more information regarding the relationship between religion and students' achievement in Arabic language, the respondents were asked about the

role that religion plays in this respect. The analyses of the interviews indicated that religion is seen to have a positive and significant relationship with Arabic learning outcomes. It was felt that many students want to learn Arabic to improve their religious knowledge, mainly understanding and memorizing Qur'an and Hadith, as stated below by the Arabic school graduate

> Ok. This relationship is significant and positive. Many students come to Linxia Arabic School because of religious improvement. Some students do not know how to pray, but they can memorize the whole Quran after graduation.

Furthermore, religious students are more motivated, more confident, and more determined to learn the Arabic language. They do not easily become discouraged to learn the language if they face any difficulties. This can be elucidated from the following statements by an Arabic school student

> I think religion is very important for learning the Arabic language. For example, it is hard to find a highly religious student complaining about his teachers, peers and school. He would try his best to overcome difficulties of learning the language, and his daily life as well. I have found that the religious awareness of good Arabic learners is stronger than of those weak learners. Many read and learn Quran to understand it and improve the Arabic language as well.

In short, according to the respondents, religion is a major motive for learning the Arabic language.

14.3.5 Suggestions on How to Improve Students' Achievement in Arabic Language

During the interviews, some suggestions were given to improve students' achievement in Arabic language. For example, students should spend more time in learning and practicing the Arabic language. They should also use the Arabic language to communicate with each other, and particularly with those who speak Arabic well, including good Arabic learners and the teachers of Arabic. In addition, parents should constantly encourage their children and provide them with every possible means to enhance their performance in Arabic. The following suggestions were given by the Arabic school graduate

> I want to give two suggestions to school students. The first is to spend more time to learn, practice, and speak Arabic; the second is to communicate with good Arabic learners and teachers of Arabic. In addition, parents should give more attention to students' learning.

RQ2 What are the relationships between the different aspects of motivation: Religious Motivation (RM), Intrinsic Motivation (IM), Extrinsic Motivation (EM), and Amotivation (AM)?

The main purpose of using Rasch analysis in this study was to produce the estimates of items and persons which are invariant and interval in nature. Table 14.4 shows item and person reliability estimates for the four aspects as well as the infit and outfit mean square fit statistics.

Table 14.4 Aspects of motivation with item and person reliability estimates

Factor	Item reliability	Items infit MNSQ	Item outfit MNSQ	Person reliability	Person infit MNSQ	Person outfit MNSQ
Overall	.98	1.01	1.00	.76	1.00	1.00
RM	.99	1.00	1.06	.67	.75	1.06
EM	.99	1.02	1.00	.66	.98	1.00
IM	.98	1.00	1.02	.72	.97	1.02
AM	.93	1.00	1.01	.69	1.00	1.01

Table 14.5 Pearson product-moment correlations between measures of aspects of motivation

		Religious	Intrinsic	Extrinsic	Amotivation
Religious	Pearson correlation	1			
	Sig. (2-tailed)				
	N	348			
Intrinsic	Pearson correlation	.420*	1		
	Sig. (2-tailed)	.000			
	N	348	348		
Extrinsic	Pearson correlation	.289*	.307*	1	
	Sig. (2-tailed)	.000	.000		
	N	348	348	348	
Amotivation	Pearson correlation	.094	.026	.005	1
	Sig. (2-tailed)	.080	.626	.926	
	N	348	348	348	348

* Correlation is significant at the 0.01 level (2-tailed)

Item reliability indices in Table 14.4 indicate high reliability index (>.90), implying that the same order of items can be expected with other comparable sample of respondents. Person reliability indices indicate acceptable though lower reliability implying limited replicability of person ordering. The mean squares for item and person infit and outfit mean are close to 1, the Rasch expected value, except for the person infit MNSQ (.75) for RM, indicating that respondents were rather more uniform in their responses on RM items than expected by the Rasch model.

To look at the relationship among the four aspects, person measure files were produced, and then, the Bivariate Pearson correlation coefficient was estimated. The results are shown in Table 14.5.

Table 14.5 shows that there was a medium positive correlation between RM and IM, $r = .420$, $n = 348$, $p < .01$, and a weak positive correlation between RM and EM, $r = 2.89$, $n = 348$, $p < .01$.A weak positive correlation was also found between EM and IM, $r = 307$, $n = 348$, $p < .01$ (see Cohen as cited in Pallant 2005). However, there was no significant correlation between AM and other aspects of motivation.

RQ3 What is the relative contribution of each aspect of motivation (Religious Motivation [RM], Intrinsic Motivation [IM], Extrinsic Motivation [EM], and Amotivation [AM]) on Arabic learning achievement?

A multiple regression analysis was conducted to answer this research question as it tells how much of the variance in the dependent variable can be explained by the

Table 14.6 Coefficients of motivation aspects

	Unstandardized coefficient		Standardized coefficient			Collinearity statistics	
Model	B	Std. error	Beta	T	Sig	Tolerance	VIF
Constant	.011	.077		.147	.883		
Religious	−.089	.040	−.131	−2.236	.026	.788	1.269
Intrinsic	−.177	.140	−.074	−1.264	.207	.786	1.272
Extrinsic	.094	.080	.065	.065	.243	.874	1.144
Amotivation	.556	.119	.243	.243	.000	.991	1.010

independent variables, and it indicates the relative contribution of each independent variable (Pallant 2005). In addition, it is important to highlight that the distributions of students' raw scores in Arabic language taken from the two schools were different because the two schools used different measurement instruments (Arabic tests) to determine the students' achievement in the Arabic language. Therefore, it would not be possible to compare the students' scores using these raw scores. To solve this issue, converting or transforming the raw scores to Z-scores in a process called 'standardization', in which the mean becomes zero and standard deviation becomes 1, makes it possible to conduct such comparison (Table 14.6).

The four variables were retained because other collinearity diagnostics were not violated as shown in Table 14.6 below, labeled as Coefficients, where two values would be the concern: Tolerance and VIF values. If the former value is less than .10, it indicates that the correlations with other variables is high, and the latter should be less than 10 to show the possibility of non-presence of multicollinearity (Pallant 2005). Table 14.6 shows that the tolerance value of each independent variable is above .10. This is supported by the VIF values which are less than 10. Therefore, multicollinearity was not violated.

Table 14.6 shows all the independent variables included in the model and how much each variable contributed to the dependent variable. The standardized coefficients were examined to compare the contribution of each independent variable. It could be seen that the largest Beta value is .243, which is for Amotivation. This means that this variable made the largest or strongest contribution on the dependent variable (i.e. Arabic language learning achievement), followed by RM, (Beta = −.131). The Beta values for IM and EM were −.074 and .064 respectively, indicating that they made the least contribution. Moreover, it is found that only the contributions of AM and RM aspects were statistically significant ($p < .05$).

In short, the multiple regression analyses show the low contribution of all the motivation aspects on Arabic language learning achievement, with significant contributions by Amotivation and Religious Motivation.

RQ4 Are there significant differences in the level of motivation (Religious Motivation [RM], Intrinsic Motivation [IM], Extrinsic Motivation [EM], and Amotivation [AM]) between *students from the Gansu Arabic Schools and the Gansu Islamic College in China*?

Table 14.7 Descriptive analysis of Arabic school and Islamic college person

Motivation aspect	Name of school	N	Mean	Std. deviation	Std. error mean
Religious	Arabic school	171	−.95	1.58	.12
	Islamic college	177	−1.02	1.36	.10
Intrinsic	Arabic school	171	−.35	.53	.04
	Islamic college	177	−.37	.28	.02
Extrinsic	Arabic school	171	−.40	.72	.05
	Islamic college	177	−.40	.66	.05
Amotivation	Arabic school	171	−.18	.48	.07
	Islamic college	177	−.25	.40	.03

Measures in terms of the various motivation aspects

The independent sample *T*-Test was conducted to find out whether there is a statistically significant difference in the mean person measures for students from Arabic school and Islamic college in each aspect of motivation. Table 14.7 shows that 348 students participated in the study (171 students from Arabic school and 177 from Islamic college).

The results of the Levene's test for equality of variances for each of the motivation aspects were checked. This test shows if the variation of scores for the two groups (Arabic school and Islamic college) is the same, and it determines which of the t-values to be used. When the significant value is larger than .05, the level of 'Equal variances assumed' should be used, and if the significant value equals or less than .05, the 'Equal variance not assumed' should be used.

Table 14.8 reveals that the significant levels for the variables RM, EM and AM were larger than .05 (.123, .695, and .741 respectively). This means that the assumption of equal variances was not violated, and so the 'Equal variance assumed' was used. Table 14.8 also shows that the significant level for the IM variable (.013) was less than .05. This means that the assumption of equal variances was violated, and so the 'Equal variance not assumed' was used. It also shows that there were no significant differences in the mean measures of the Arabic school and Islamic college on all aspects of motivation. The p values (two-tailed) of all the aspects were greater than .05 (RM, $p = .653$; IM, $p = .664$; EM, $p = .179$; and AM, $p = 149$).

14.4 Discussion

This study primarily aims to explore factors that affect Arabic learning achievement of Chinese students from Gansu Islamic School, China and examine the relative contribution of the different aspects of motivation Intrinsic Motivation (IM), Extrinsic Motivation (EM), Amotivation (AM), and Religious motivation (RM) on their Arabic learning achievement. It also investigates the relationship between the four aspects of motivation and compares their levels between students from two types of Gansu Islamic Schools in China.

Table 14.8 Independent sample *T*-Test

Variable		Levene's test for equality of variances		t	Df	Sig. (2-tailed)	Mean difference	Std. error difference	95 % confidence interval of the difference	
		F	Sig.							
RM	Equal variances assumed	2.391	.123	.450	346	.653	.0711	.15803	−.23972	.38191
	Equal variances not assumed			.449	334.438	.654	.0711	.15844	−.24058	.38276
IM	Equal variances assumed	6.201	.013	.439	346	.661	.0197	.04486	−.06853	.10792
	Equal variances not assumed			.435	252.972	.664	.0197	.04531	−.06953	.10893
EM	Equal variances assumed	.154	.695	1.347	346	.179	.0995	.07387	−.04578	.24482
	Equal variances not assumed			1.345	340.964	.179	.0995	.07399	−.04600	.24505
AM	Equal variances assumed	.109	.741	1.446	346	.149	.0676	.04678	−.02438	.15964
	Equal variances not assumed			1.441	328.088	.151	.0676	.04694	−.02472	.15998

The first research question asked what factors contribute to students' Arabic leaning achievement from the perspectives of seven respondents (a rector of an Arabic-school, a teacher of the Arabic language, a Chinese Arabic-school graduate student, and four school students learning Arabic language), through face-to-face interviews. The qualitative analyses of the interview data identified some internal and external factors that the respondents felt could significantly contribute to achievement in the Arabic language. Both these internal and external factors are part of 'intrinsic motivation' and 'extrinsic motivation' proposed by many researchers (Deci and Ryan 1985, 2000; Dornyei 1998a, b; Gardner 1985).

With respect to the relationship among the four aspects of motivation, it was found that there is a positive relationship between three aspects of motivation (Religious Motivation [RM], Intrinsic Motivation [IM], and Extrinsic Motivation [EM]). This could be explained via some studies (Deci 1971; Deci and Ryan 1985; Gibbons 1997; Lazear 2000). This study also found that there was no significant correlation between amotivation (AM) and other aspects of motivation. SDT researchers highlighted that amotivation comes from internal and external sources that lead individuals to lose interest in performing tasks or achieving a goal (Kasser and Ryan 1993, 1996; McHoskey 1999; Sheldon and McGregor 2000; Sheldon et al. 2000).

The multiple regression analysis showed that AM and RM had the largest and significant contribution on Arabic language learning achievement. AM contributed negatively towards Arabic language learning whereas RM contributed positively. As discussed earlier, some studies showed a strong relationship between religion and language learning. Palmer (2007) found religion could motivate learners to acquire language; Jaspal and Coyle (2009) indicated religion plays an important role in students' Arabic language learning. These findings are consistent with Bakar et al. (2010)'s. This study also revealed Amotivationas a significant negative influence on students' Arabic learning outcomes, and further showed a slight and statistically non-significant contribution of intrinsic motivation and extrinsic motivation on Arabic learning achievement.

Finally, the independent sample *T*-Test showed that there were no significant differences in the level of motivation between students of the Arabic school and Islamic college on all aspects of motivation. This means that regardless of the school type, students may share the same attitudes towards the different aspects of motivation. And it is possible that these schools have almost the same facilities and learning conditions.

References

Bakar, K. A., Sulaiman, N. F., & Rafaai, Z. A. M. (2010). Self-determination theory and motivational orientations of Arabic learners: A principal component analysis. *Journal of Language Studies, 10*(1), 71–86.

Chambers, G. N. (1993). Talking the 'de' out of demotivation. *Language Learning Journal, 7*, 13–16.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*, 105–115.

Deci, E. L. (1980). *The psychology of self-determination*. Lexington: Health.
Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self determination in human behavior*. New York: Plenum Press.
Deci, E. L., & Ryan, R. M. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*, 54–67.
Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology, 49*(3), 182–185.
Dornyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal, 78*(3), 273–284.
Dornyei, Z. (1996). Moving language learning motivation to a larger platform for theory and practice. In R. L. Oxford (Ed.), *Language learning motivation: pathways to the New Century* (pp. 71–80). Honolulu: University of Hawaii Press.
Dornyei, Z. (1998a). *Demotivation in foreign language learning*. Paper presented at the TESOL'98 Congress, Seattle.
Dornyei, Z. (1998b). Motivation in second and foreign language teaching. *Language Teaching, 31* (3), 117–135.
Dornyei, Z. (2003). Attitudes, orientation, and motivations in language learning: Advances in theory, research, and applications. *Language Learning, 53* (Suppl), 3–32.
Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. Baltimore: Edward Arnold.
Gardner, R. C. (2001). Integrative motivation and second language acquisition. In Z. Dornyei & R. Schmidt (Eds.), *Motivation and second language learning* (pp. 1–20). Honolulu: University of Hawai 'i Press.
Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second language learning*. Rowley: Newbury House Publishers, Inc.
Gibbons, R. (1997). Incentives and careers in organizations. In D. Kreps & K. Wallis (Eds.), *Advances in economic theory and econometrics* (Vol. II). Cambridge: Cambridge University Press.
Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: Motivational and informational components. *Developmental Psychology, 17*, 300–312.
Jaspal, R., & Coyle, A. (2009). Language and perceptions of identity threat. *Psychology and Society, 2*(2), 150–167.
Kasser, T., & Ryan, R. M. (1993). A dark side of the American dream: Correlates of financial success as a central life aspiration. *Journal of Personality and Social Psychology, 65*, 410–422.
Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin, 22*, 280287.
Keblawi, F. (2009). A review of language learning motivation theories. 23–57. Retrieved from http://www.qsm.ac.il/mrakez/asdarat/jamiea/12/eng-2-faris%20Keblawi.pdf. Accessed 28 Nov 2010.
Lazear, E. (2000). Performance, pay and productivity. *The American Economic Review, 90*(5), 1346–1361.
Li, X. H., & Feng, J. Y. (1998). *History of Islam in China*. China Social Science Press. Retrieved from http://contemporary_chinese_culture.academic.ru/370/Islam_in_China. Accessed 11 Dec 2010.
Lucas, R. I., Pulido, D., Miraflores, E., Ignacio, A., Tacay, M., & Lao, J. (2010). A study on the intrinsic motivation factors in second language learning among selected freshman students. *Philippine ESL Journal, 4*, 3–22.
MacIntyre, P. D., MacMaster, K., & Baker, S. C. (2001). The convergence of multiple models of motivation for second language learning: Gardner, Pintrich, Kuhl, and McCroskey. In Z. Dornyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (Technical Report #23, pp. 461–492). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.
Malik, T. G. (2010). Role of L2 Motivation and the performance of intermediate students in the English (L2) exams in Pakistan. *Language in India, 10*(7), 37–49.

McHoskey, J. W. (1999). Machiavellianism, intrinsic versus extrinsic goals, and social interest: A self-determination theory analysis. *Motivation and Emotion, 23*, 267–283.

Noels, K. A. (2001a). Learning Spanish as a second language: Learners' orientations and perceptions of their teachers' communication style. *Language Learning, 51*(1), 107–144.

Noels, K. A. (2001b). New orientations in language learning motivation: Towards a model of intrinsic, extrinsic and integrative orientations. In Z. Dornyei & R. W. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 43–68). Honolulu: University of Hawai'i Second Language Teaching and Curriculum Centre.

Noels, K. A., Clement, R., & Pelletier, L. G. (1999). Perceptions of teachers' communicative style and students' intrinsic and extrinsic motivation. *The Modern Language Journal, 83*, 23–34.

Noels, K. A., Pelletier, L. G., Clement, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self determination theory. *Language Learning, 50*, 57–85.

Noels, K. A., Clement, R., & Pelletier, L. G. (2001). Intrinsic, extrinsic, and integrative orientations of French Canadian learners of English. *The Canadian Modern Language Review, 57*, 424–444.

Oxford, R. L. (1998). Anxiety and the language learner: New insights. In J. Arnold (Ed.), *Affective language learning*. Cambridge: Cambridge University Press.

Oxford, R. L., & Shearin, J. (1994). Language learning motivation: Expanding the theoretical framework. *The Modern Language Journal, 1*(78), 12–28.

Pallant, J. (2005). *SPSS survival guide,* (2nd ed.). Sydney: Open University Press.

Palmer, J. (2007). Arabic diglossia: Teaching only the standard variety is a disservice to students. *Arizona Working Papers in SLA & Teaching, 14*, 111–122.

Ryan, R. M., & Stiller, J. (1991). The social contexts of internalization: Parent and teacher influences on autonomy, motivation and learning. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement* (Vol. 7, pp. 115–149). Greenwich: JAI Press.

Sheldon, K. M., & McGregor, H. (2000). Extrinsic value orientation and the tragedy of the commons. *Journal of Personality, 68*, 383–411.

Sheldon, K. M., Sheldon, M. S., & Osbaldiston, R. (2000). Prosocial values and group-assortation within an N-person prisoner's dilemma. *Human Nature, 11*, 387–404.

Ushioda, E. (1998). Effective motivational thinking: A cognitive theoretical approach to the study of language learning motivation. In E. A. Soler & V. C. Espurz (Eds.), *Current issues in English language methodology* (pp. 77–89). Castello de la Plana: Universitat Jaume I.

Vallerand, R. J., Blais, M. R., Briere, N. M., & Pelletier, L. G. (1989). Construction and validation of the motivation toward education scale. *Canadian Journal of Behavioral Science Revue Canadienne, 21*, 323–349.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The academic motivation scale: A measurement of intrinsic, extrinsic and amotivation in education. *Educational and Psychological Measurement, 52*, 1003–1016.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Senecal, C., & Vallieres, E. F. (1993). On the assessment of intrinsic, extrinsic and amotivation in education: Evidence on the concurrent and construct validity of the academic motivation scale. *Educational and Psychological Measurement, 53*, 159–172.

Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology, 72*(5), 1161–1176.

Chapter 15
Rasch-Based Analysis of Item and Person Fit – A Language Testing Practice in Jiaxing University China

Guoxiong He and Huifeng Mu

Abstract The study is to analyze item and person fit to the Rasch model, aiming to improve test quality and facilitate item banking. Five-hundred 1st-year non-English major students from Jiaxing University are randomly sampled in the study. The scores were obtained from a reading subtest, which consists of 30 multiple-choice items. A software called GITEST III + SYSTEM was used to calculate item difficulty and candidate ability. For each item and each test taker, Rasch model fit statistics were also calculated, including Wright's squared standardized residuals (W) and associated t values. Other indices such as Bock's chi-square (BCHI) and Drasgow's standardized appropriateness index (l_z) are used for further illustration. The results from Wright's W show that six items and 40 candidates are identified as misfitting, but BCHI and l_z produce the outcomes with different number of misfitting items and candidates, suggesting that several procedures are needed to check the degree of misfit.

Keywords Rasch • Fit analysis • Item difficulty

15.1 Background

Person-fit statistics have been proposed to investigate the fit of an item score pattern to an item response theory (IRT) model. In language testing or any other testing, a candidate who can pass an item of some difficulty will be able to answer any less difficult item correctly, and if he fails an item with certain difficulty, it implies that he can't pass any item more difficult than that. In other words, they should get items correct that are below their ability and should get items wrong that are above their ability. We can use a dichotomous response matrix called implicational scale to

G. He (✉) • H. Mu
Faculty of Foreign Studies, Jiaxing University, Jiaxing, China
e-mail: hgx505@mail.zjxu.edu.cn

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_15,

Table 15.1 Implicational table

Student no.	Item 1	Item 2	Item 3	Item 4	Item 5
1.	0	0	0	0	0
2.	1	0	0	0	0
3.	1	1	0	0	0
4.	1	1	1	0	0
5.	1	1	1	1	0
6.	1	1	1	1	1

From Davies et al. (2002: 80)

illustrate that (see the following table). Because it was first used by L. Guttman, (Xu, 1992: 32); it is also called Guttman perfect scale. From the table, we can see any subject who gives the correct answer to Item 5 will know the answers to all other items in the set. On the other hand, those who supply a wrong response to Item 3 will be unable to do Items 4 and 5 (Davies et al., 2002: 80).

But in practice, this is not always the case (Du, 2007: 71). For some reason or other, learners will sometimes produce unexpected answers. Some candidates answer a few more difficult test items correctly, but give wrong answers to some easier ones, thus abnormal response patterns occur. Suppose a test taker gives answers to five items with a response pattern of 00111, he gets the easiest two items wrong but the difficult three right, which is not what we expect. Usually we interpret scores from tests as indicators of ability, but when scores are abnormal or unusual, they are not good indicators of candidate's performance, and we have no reason to claim the test taker's test score is a valid measure of his ability. This brings us into the analysis of abnormal response patterns on language tests to aim for high quality testing.

15.2 Research Design

15.2.1 Research Questions

The study is to analyze item and person fit to the Rasch model. What we test is whether the Rasch model could provide an adequate fit or bad fit to the test data observed in a random sample, or alternatively weather the Rasch model adequately represents the testing data. Specifically, the following questions are addressed:

1. How many items exhibited unlikely response patterns and are rejected as misfitting?
2. How many candidates exhibited improbable response patterns and are identified as misfitting?

15.2.2 Subjects

Five-hundred 1st-year non-English major students from Jiaxing University, 18 years of age and up, male and female, are randomly sampled in the study.

They took a reading course in College English, and took the final examination at the end of the first semester.

15.2.3 Procedure

As the study was aimed at analyzing item and person fit to the Rasch model, the scores were obtained from a reading subtest on a final examination, consisting of 30 multiple-choice items, namely, dichotomously scored test items.

The scores from 500 candidates on 30 items were arranged in successive order of difficulty and ability. GITEST III + SYSTEM (Gui, 1991: 11–58; Zhang, 2004:16), which is China's first testing analysis system developed by Guangdong Foreign Studies University in 1983 and widely used among more than 80 colleges and universities in China, was used to calculate item difficulty and candidate ability. For each item and each test taker, Rasch model fit statistics were also calculated, including the sum of squared standardized residuals and associated t values. The procedure is usually referred to as squared standardized residual which was put forward by Wright in 1977. The establishment of a critical value of t for identifying misfitting items and candidates is arbitrary, but commonly the t-value of positive 2.00 is set as the critical value, so that items or candidates with t-value of positive 2.00 or above are considered misfits to the model and lack response validity (Henning, 2001: 123). For each item, point biserial correlation is also calculated to compare with the corresponding fit statistic.

In the meantime, we split up the candidates into six groups on the basis of average ability, each group being roughly considered homogenous subpopulation. A chi-square fit statistic called BCHI was used in order to check weather the observed data can fit the IRT one parameter, or rather, how many misfitting items can be found out according to the chi-square test.

We also take the natural logarithm of the likelihood function at the max. As this is dependent on the level of theta, so we need it standardized. This is called l_z, like a z-score, less than -2 means bad fit.

15.3 Results and Interpretation

The results given by GITEST III + SYSTEM show that six items were identified as misfitting by the conventional criterion of fit t surpassing 2.00.

We see from Table 15.2 that the six items with t-value surpassing 2.00 are identified: Items 2,3,5,7,12 and 13. We also see that the point biserial correlation coefficients (Qi, Dai & Ding 2002: 111) between each of the six item scores and subtest scores are all rather low, which tells us that the six items are quite different from the rest of the items. One parameter model assumes all the items in the test are of equal discrimination. If the point biserial correlation is too low or too high, the observed response patterns might not conform to the model.

Table 15.2 Item and person Rasch model fit statistics

Item	n corr	ln(Q/P)	ini dif	fi dif	s.e.	Σz^2	t	pbi
1.	347	−0.82	0.36	0.41	0.10	550.35	1.59	0.21
2.	320	−0.58	0.61	0.68	0.10	599.92	3.05*	0.17
3.	363	−0.97	0.21	0.23	0.11	583.96	2.59*	0.19
4.	318	−0.56	0.62	0.70	0.10	489.28	−0.31	0.36
5.	204	0.37	1.55	1.75	0.10	566.26	2.06*	0.26
6.	277	−0.22	0.97	1.09	0.10	512.07	0.41	0.36
7.	228	0.18	1.36	1.53	0.10	602.14	3.12*	0.23
8.	444	−2.07	−0.89	−1.00	0.15	398.68	−3.36	0.37
9.	420	−1.66	−0.48	−0.54	0.13	437.60	−2.01	0.33
10.	435	−1.90	−0.72	−0.81	0.14	343.33	−5.42	0.45
11.	309	−0.48	0.70	0.79	0.10	477.56	−0.69	0.38
12.	107	1.30	2.48	2.80	0.12	759.94	7.45*	0.13
13.	257	−0.06	1.13	1.27	0.10	611.16	3.38*	0.17
14.	434	−1.88	−0.70	−0.79	0.14	411.73	−2.90	0.40
15.	358	−0.92	0.26	0.29	0.11	462.24	−1.19	0.40
16.	484	−3.41	−2.23	−2.51	0.27	381.19	−3.99	0.27
17.	388	−1.24	−0.06	−0.07	0.11	524.95	0.81	0.31
18.	420	−1.66	−0.48	−0.54	0.13	391.72	−3.61	0.40
19.	448	−2.15	−0.97	−1.10	0.16	409.45	−2.98	0.39
20.	484	−3.41	−2.23	−2.51	0.27	239.87	−9.89	0.40
21.	433	−1.87	−0.68	−0.77	0.14	461.66	−1.21	0.37
22.	396	−1.34	−0.15	−0.17	0.12	510.35	0.36	0.27
23.	476	−2.99	−1.81	−2.04	0.22	262.97	−8.79	0.42
24.	285	−0.28	0.90	1.02	0.10	464.72	−1.10	0.40
25.	374	−1.09	0.09	0.11	0.11	407.13	−3.06	0.48
26.	291	−0.33	0.85	0.96	0.10	501.81	0.09	0.35
27.	457	−2.36	−1.18	−1.33	0.17	292.65	−7.48	0.48
28.	316	−0.54	0.64	0.72	0.10	513.92	0.47	0.36
29.	353	−0.88	0.31	0.35	0.10	430.17	−2.26	0.46
30.	420	−1.66	−0.48	−0.54	0.13	360.20	−4.77	0.47

*Means the observed data do not match the Rasch Model

It is very difficult to use one single statistic to check the fit of the model to the data (Yu, 1992: 183). To decide whether an item fits the model, some other criterion needs to be employed (Croker & Algina, 2004: 406). For illustrative purposes, we carried out a chi-square test. Let's look at Table 15.3, which is the observed proportion correct based on the response data. The candidates are divided into six groups with the first group having the lowest mean ability (−0.51) and the sixth group the highest (2.71). The data in the table show the proportion correct of each group on each item. For example, for Item 1, 62 % of candidates from Group 1 got it right, 57 % from Group 2 got it right, and so on. When the mean ability of the group increases, the proportion correct should also increase. If the data fit the model, they should be in agreement with the expected percentage or the expected probability of ICC (Table 15.4). If the data don't conform to Rasch model, some difference would exist.

Table 15.3 Observed proportion correct

Item	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
1.	0.62	0.57	0.61	0.62	0.86	0.86
2.	0.51	0.57	0.54	0.62	0.77	0.76
3.	0.49	0.72	0.68	0.74	0.73	0.89
4.	0.24	0.41	0.60	0.73	0.71	0.86
5.	0.16	0.27	0.38	0.38	0.46	0.68
6.	0.27	0.26	0.48	0.59	0.67	0.88
7.	0.30	0.30	0.36	0.47	0.55	0.68
8.	0.51	0.80	0.92	0.91	0.96	0.99
9.	0.49	0.70	0.83	0.89	0.92	0.99
10.	0.46	0.69	0.86	0.95	0.99	0.99
11.	0.30	0.38	0.51	0.67	0.76	0.89
12.	0.19	0.16	0.16	0.14	0.30	0.35
13.	0.38	0.51	0.41	0.50	0.48	0.79
14.	0.46	0.74	0.91	0.91	0.93	0.99
15.	0.24	0.55	0.66	0.77	0.85	0.92
16.	0.81	0.96	0.96	0.99	0.99	1.00
17.	0.38	0.74	0.73	0.83	0.83	0.90
18.	0.46	0.70	0.84	0.83	0.97	1.00
19.	0.57	0.77	0.90	0.97	0.95	0.99
20.	0.70	0.99	0.97	0.99	1.00	1.00
21.	0.49	0.76	0.90	0.90	0.95	0.94
22.	0.49	0.73	0.77	0.81	0.84	0.96
23.	0.65	0.92	0.97	0.98	1.00	1.00
24.	0.22	0.34	0.40	0.59	0.74	0.94
25.	0.27	0.46	0.72	0.81	0.93	0.96
26.	0.22	0.39	0.52	0.56	0.76	0.83
27.	0.46	0.81	0.96	0.97	0.98	1.00
28.	0.24	0.45	0.55	0.70	0.74	0.88
29.	0.16	0.47	0.68	0.76	0.88	0.93
30.	0.38	0.70	0.80	0.90	0.96	1.00
n	37	74	98	115	104	72
Range	4–15	16–19	20–21	22–23	24–25	26–30
Mean b	−0.51	0.47	0.97	1.36	1.84	2.71

Table 15.5 shows the differences between the observed percentages and the expected percentages. Take Item 13 for example, 38 % from group 1 got the item correct, while the expected value is 14 %, so the difference is 24 %. Similarly, the difference from group 2 on that item is 20 % (51–31 %). Obviously, it does not conform to the ICC. The result of chi-square test (Table 15.6) show that the chi-square value of item 13 is 42.36, much larger than the corresponding critical value of 15.07 ($df = 5$, $p < 0.01$).

From Table 15.6 we see that 7 other items besides item 13 are found out to have the chi-square values exceeding the critical value of 15.07. Others are items 1, 2, 3, 7, 12, 20 and 27. But items 1, 20 and 27 were not identified as misfitting according to t-values. Although different test procedures do not give exactly the same result, the intersection of the two procedures contains 5 items: items 2, 3, 7, 12

Table 15.4 Expected proportion

Item	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
1.	0.28	0.52	0.64	0.72	0.81	0.91
2.	0.23	0.45	0.57	0.66	0.76	0.88
3.	0.32	0.56	0.67	0.76	0.83	0.92
4.	0.23	0.44	0.56	0.66	0.76	0.88
5.	0.09	0.22	0.31	0.40	0.52	0.72
6.	0.17	0.35	0.47	0.57	0.68	0.84
7.	0.11	0.26	0.36	0.46	0.58	0.76
8.	0.62	0.81	0.88	0.91	0.94	0.98
9.	0.51	0.73	0.82	0.87	0.92	0.96
10.	0.57	0.78	0.86	0.90	0.93	0.97
11.	0.21	0.42	0.54	0.64	0.74	0.87
12.	0.04	0.09	0.14	0.19	0.28	0.48
13.	0.14	0.31	0.42	0.52	0.64	0.81
14.	0.57	0.78	0.85	0.90	0.93	0.97
15.	0.31	0.55	0.66	0.74	0.82	0.92
16.	0.88	0.95	0.97	0.98	0.99	0.99
17.	0.39	0.63	0.74	0.81	0.87	0.94
18.	0.51	0.73	0.82	0.87	0.92	0.96
19.	0.64	0.83	0.89	0.92	0.95	0.98
20.	0.88	0.95	0.97	0.98	0.99	0.99
21.	0.56	0.78	0.85	0.89	0.93	0.97
22.	0.42	0.66	0.76	0.82	0.88	0.95
23.	0.82	0.92	0.95	0.97	0.98	0.99
24.	0.18	0.37	0.49	0.59	0.70	0.84
25.	0.35	0.59	0.70	0.78	0.85	0.93
26.	0.19	0.38	0.50	0.60	0.71	0.85
27.	0.69	0.86	0.91	0.94	0.96	0.98
28.	0.23	0.44	0.56	0.65	0.75	0.88
29.	0.30	0.53	0.65	0.73	0.82	0.91
30.	0.51	0.73	0.82	0.87	0.92	0.96

and 13. It is reasonable to consider the 5 items as misfits, and the other items should be interpreted with caution.

As for person misfits, GITEST III + SYSTEM identified 40 with the t-values exceeding the critical value 2.00 (Table 15.7). Take person 40 for example, a low ability examinee who only got 4 items right, with an ability estimated at only −2.31, succeeded with a high difficulty item, with a difficulty of 1.27. It was possible that the candidate was blindly guessing. On the other hand, candidate 4 with an ability of 1.99, missed item 16, the easiest item of all, with a difficulty of −2.51. Obviously, their response patterns are abnormal or unusual. Scores obtained by abnormal test takers do not have the same meaning as the corresponding scores by normal candidates (Hulin, Drasgow & Parsons, 1990: 165).

There are several procedures to check abnormal responses. We also calculated the l_z index put forward by Drasgow. What is different from the procedures above is

Table 15.5 Difference from expected ICC

Item	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
1.	0.34	0.05	−0.02	−0.10	0.05	−0.05
2.	0.28	0.12	−0.03	−0.05	0.01	−0.12
3.	0.16	0.16	0.01	−0.02	−0.10	−0.03
4.	0.01	−0.04	0.04	0.07	−0.05	−0.02
5.	0.07	0.05	0.07	−0.02	−0.06	−0.04
6.	0.10	−0.09	0.01	0.02	−0.01	0.04
7.	0.18	0.04	0.00	0.01	−0.03	−0.08
8.	−0.11	−0.02	0.04	0.00	0.02	0.01
9.	−0.02	−0.03	0.01	0.02	0.01	0.02
10.	−0.12	−0.09	0.00	0.05	0.06	0.01
11.	0.08	−0.04	−0.03	0.03	0.02	0.02
12.	0.15	0.07	0.03	−0.05	0.02	−0.13
13.	0.23	0.20	−0.02	−0.02	−0.16	−0.02
14.	−0.11	−0.04	0.06	0.02	0.00	0.02
15.	−0.07	0.01	0.00	0.03	0.02	0.00
16.	−0.07	0.01	−0.01	0.01	0.00	0.01
17.	−0.01	0.11	0.00	0.03	−0.04	−0.04
18.	−0.05	−0.03	0.02	−0.03	0.06	0.04
19.	−0.07	−0.06	0.01	0.05	0.00	0.01
20.	−0.18	0.03	0.00	0.01	0.01	0.01
21.	−0.08	−0.02	0.05	0.01	0.02	−0.03
22.	0.07	0.07	0.01	−0.01	−0.05	0.01
23.	−0.17	−0.01	0.02	0.02	0.02	0.01
24.	0.04	−0.03	−0.09	0.01	0.05	0.10
25.	−0.08	−0.13	0.02	0.03	0.08	0.03
26.	0.03	0.01	0.02	−0.04	0.05	−0.02
27.	−0.24	−0.05	0.05	0.04	0.02	0.02
28.	0.02	0.01	−0.01	0.04	−0.01	0.00
29.	−0.14	−0.06	0.03	0.02	0.06	0.02
30.	−0.13	−0.03	−0.02	0.03	0.05	0.04

that the value of l_z less than -2 is considered to be bad fit. The smaller l_z is, the more abnormal the response pattern will be. In the light of this criterion, 19 candidates are found out to be misfitting, 16 of whom are also considered as misfitting by the squared standardized residual. But to some degree, the critical value of -2 is somewhat arbitrary. If the critical value is -0.5, then 104 misfits are found out, and the intersection of two procedures contains 33 misfitting candidates.

15.4 Discussion

As one of the testing purposes is to assess candidate ability, which is usually embodied in the test scores, a proper assessment seems absolutely vital. However, some high ability candidates may answer several easier items incorrectly, while some low ability examinees can give correct responses to a few more difficult items. These misfitting

Table 15.6 Bock's chi-square

Item	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Chi-square
1.	20.60	0.81	0.23	6.17	1.60	1.99	31.39
2.	16.41	4.36	0.33	1.06	0.04	10.02	32.22
3.	4.59	7.43	0.03	0.16	7.76	1.14	21.11
4.	0.04	0.40	0.55	2.65	1.16	0.29	5.10
5.	2.01	1.23	1.93	0.20	1.50	0.63	7.50
6.	2.77	2.85	0.04	0.27	0.02	0.83	6.78
7.	12.14	0.63	0.01	0.07	0.34	2.83	16.02
8.	1.79	0.13	1.54	0.00	0.55	0.31	4.32
9.	0.06	0.34	0.05	0.31	0.09	1.11	1.95
10.	2.00	3.81	0.00	3.15	5.36	0.57	14.90
11.	1.53	0.55	0.44	0.48	0.20	0.18	3.37
12.	25.86	4.94	0.55	2.03	0.24	4.90	38.52
13.	16.45	14.29	0.10	0.15	11.23	0.13	42.36
14.	1.83	0.56	2.40	0.37	0.00	0.60	5.76
15.	0.77	0.02	0.00	0.52	0.33	0.00	1.64
16.	1.74	0.09	0.40	0.79	0.08	0.39	3.48
17.	0.03	3.96	0.00	0.58	1.77	1.96	8.31
18.	0.33	0.34	0.23	1.23	4.21	2.80	9.14
19.	0.90	1.69	0.12	4.43	0.01	0.21	7.35
20.	11.21	1.94	0.00	0.79	1.34	0.39	15.66
21.	0.92	0.16	1.75	0.13	0.68	1.65	5.29
22.	0.74	1.78	0.03	0.16	2.10	0.18	4.98
23.	7.52	0.03	0.61	0.83	2.16	0.62	11.77
24.	0.36	0.27	3.13	0.02	1.01	5.44	10.23
25.	1.04	5.23	0.23	0.63	5.59	0.83	13.56
26.	0.21	0.04	0.15	0.85	1.41	0.20	2.86
27.	9.64	1.39	3.01	2.70	1.19	1.26	19.19
28.	0.07	0.02	0.03	0.88	0.09	0.01	1.11
29.	3.27	1.01	0.49	0.30	2.36	0.25	7.68
30.	2.43	0.34	0.32	1.23	2.89	2.80	10.00

responses mean that the responses to the test items are inconsistent, suggesting a pattern of carelessness or inattention, guessing or cheating, or that the candidate abilities are not appropriately measured by the test, to the effect that the items are poorly designed, or are measuring something different from the rest of the test.

A person's response pattern reflects deviation from expected model. We need some type of fit statistic to check the model data fit. But no single statistical test can fully interpret item and person fit to the model because different methods produce different results. Several various procedures are needed to explain the degree of misfit, the unlikelihood of success or failure in responding to a test item.

Besides, fit statistics are also dependent on the sample size (Yu, 1992: 181–182). If a sample is large enough, misfitting items are quickly found out. On the other hand, if a sample is quite small, even if misfitting items do exist, it is still hard to identify them. Hambleton, R.K. and Rovinelli, R. studied the relationship between

Table 15.7 Misfitting persons

Student	Score	Initial measure	Final measure	Standard error	z^2	t	Lz
1.	27	2.20	2.71	0.68	54.32	2.86	−0.46
2.	27	2.20	2.71	0.68	56.39	3.06	−0.29
3.	26	1.87	2.31	0.60	49.53	2.37	−0.77
4.	25	1.61	1.99	0.54	120.24	8.70	−1.74
5.	24	1.39	1.71	0.51	47.18	2.12	−1.21
6.	23	1.19	1.47	0.48	50.53	2.47	−0.26
7.	22	1.01	1.25	0.46	63.75	3.78	−0.34
8.	22	1.01	1.25	0.46	66.69	4.06	−1.09
9.	22	1.01	1.25	0.46	48.69	2.28	−0.76
10.	21	0.85	1.05	0.44	55.50	2.98	−0.43
11.	21	0.85	1.05	0.44	57.10	3.13	−0.88
12.	21	0.85	1.05	0.44	57.63	3.19	−0.81
13.	21	0.85	1.05	0.44	50.64	2.48	−1.37
14.	21	0.85	1.05	0.44	61.38	3.55	−1.01
15.	20	0.69	0.86	0.43	56.20	3.05	−1.29
16.	20	0.69	0.86	0.43	48.86	2.30	−0.32
17.	20	0.69	0.86	0.43	48.93	2.30	−0.22
18.	20	0.69	0.86	0.43	47.37	2.14	−1.37
19.	19	0.55	0.68	0.42	56.88	3.11	−1.78
20.	19	0.55	0.68	0.42	47.57	2.16	−2.01
21.	19	0.55	0.68	0.42	52.92	2.72	−2.35
22.	19	0.55	0.68	0.42	53.51	2.78	−1.28
23.	19	0.55	0.68	0.42	51.31	2.55	−3.04
24.	19	0.55	0.68	0.42	55.53	2.98	−1.98
25.	17	0.27	0.33	0.41	66.93	4.08	−4.00
26.	17	0.27	0.33	0.41	49.47	2.36	−2.92
27.	14	−0.13	−0.16	0.41	48.62	2.27	−1.41
28.	13	−0.27	−0.33	0.41	61.59	3.57	−2.39
29.	13	−0.27	−0.33	0.41	52.08	2.63	−2.74
30.	13	−0.27	−0.33	0.41	68.66	4.24	−3.64
31.	13	−0.27	−0.33	0.41	48.82	2.29	−2.91
32.	11	−0.55	−0.68	0.42	59.94	3.41	−3.59
33.	10	−0.69	−0.86	0.43	48.03	2.21	−2.41
34.	9	−0.85	−1.05	0.44	79.53	5.24	−1.75
35.	9	−0.85	−1.05	0.44	60.52	3.47	−3.15
36.	9	−0.85	−1.05	0.44	101.49	7.14	−3.75
37.	9	−0.85	−1.05	0.44	85.60	5.78	−2.62
38.	9	−0.85	−1.05	0.44	99.79	7.00	−3.53
39.	6	−1.39	−1.71	0.51	77.81	5.08	−2.83
40.	4	−1.87	−2.31	0.60	64.09	3.81	−1.40

sample sizes and t statistics, with the candidates from 150 to 2,400, the number of misfitting items increased from 20 to 42 (50 items in total). Hence the t statistic is very sensitive to changes in sample size. That's why we need several procedures for checking the degree of misfit.

Fit analysis is a relatively new research area in testing theory (Xu, 1992: 122). The various fit statistics have up to now not widely employed in most examinations, but it would be helpful to make rational decisions on candidate ability and item quality, to establish a criterion for accepting or rejecting items, as well as to develop an item bank. Therefore, the importance of fit analysis should not be underestimated.

Appendix

References

Croker, L., & Algina, J. (2004). *Introduction to classical and modern test theory [M]* (trans: Jin Yu et al.). Shanghai: Eastern China Normal University Press.

Davies, A., et al. (2002). *Dictionary of language testing*. Beijing: Foreign Language Teaching and Research Press.

Du Wenjiu. (2007). *Advanced item response theory*. Chongqing: Southwest Normal University Press.

Gui Shichun. (1991). *Item bank building [A]*. Beijing: Bright Daily Press. Education Counseling Center. (1990). *Item response theory – Application in psychometrics [C]* (pp. 11–58). Beijing: Guangming Daily Press.

Henning, G. (2001). *A guide to language testing: Development, evaluation and research*. Beijing: Foreign Language Teaching and Research Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1990). *Item Response Theory — Application to Psychological Measurement [M]* (trans: East China Normal University). Wuhan: Hubei Education Press.

Qi Shuqing, Dai Haiqi, & Ding Shuliang. (2002). *Principle of modern education and psychometrics*. Beijing: Higher Education Press.

Xu Zuwei. (1992). *Item response theory and its application in testing*. Shanghai: Eastern China Normal University Press.

Yu Jiayuan. (1992). *Item response theory and its application [M]*. Nanjing: Jiangsu Education Press.

Quan Zhang. (2004). *Item analysis and equating in language testing: Research & application*. Beijing: Higher Education Press.

Chapter 16
The Contribution of Lower-Level Processing to Foreign Language Reading Comprehension with Chinese EFL Learners

Feifei Han

Abstract This paper reports an ongoing research project, which investigates the two competing hypotheses: whether inefficient lexical access (LA) and small working memory (WM) inhibit text comprehension in FL reading (inhibition hypothesis) or whether readers could use strategies to compensate for processing and language problems so that text comprehension are not influenced much (compensation hypothesis). Four hundred and two Chinese university students in their second year participated in the study. The larger project adopted a mix-method design collecting both quantitative and qualitative data, but the focus of the presentation is only on the quantitative data. The reading text were analyzed using RUMM2030 for both checking both misfitting items and persons. The data were then analyzed by doing correlation analyses between LA, WM and reading comprehension in two reading conditions: untimed reading and timed reading. Firstly, there was no significant correlation between students' LA and reading comprehension in untimed reading condition ($r = -.09, p = .07$), whereas a small and negative relationship was found between LA and reading comprehension in timed reading ($r = -.22$, $p < .01$). Secondly, WM showed to be correlated positively with students' reading comprehension in both untimed ($r = .11$, $p < .05$) and timed reading conditions ($r = .20$, $p < .01$), both the values of correlation were small. The magnitude of correlation between WM and comprehension in untimed reading was smaller than that between WM and comprehension in timed reading. The preliminary results from the quantitative data seem to support the compensation hypothesis that when readers are allowed sufficient time, inefficient word processing and small working memory do not inhibit text comprehension in FL reading.

F. Han (✉)
Faculty of Education and Social Work, The University of Sydney, Camperdown, NSW, Australia
e-mail: feifei.han@sydney.edu.au

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_16,

16.1 Introduction and Literature Review

The primary goal of reading is to construct a mental representation of meaning from a text (Grabe and Stoller 2002). This meaning constructing activity involves coordination of multiple levels of sub-component processes, including lower-level processes and higher-level processes (Grabe 2009). Theoretical models of reading place different emphases on the roles of lower-level processing and higher-level processing in reading comprehension. On one side of the coin, some researchers stress the importance of efficiency of lower-level processing, suggesting that inefficiency in word processing inhibits higher-level reading comprehension processes, such as uses of reading strategies and text comprehension (e.g. Perfetti 1988; Perfetti and Hart 2001). In the present study, this hypothesis is referred to as the inhibition hypothesis. On the other side of the coin, other researchers emphasize the role of strategic processing in reading comprehension, maintaining that, as long as readers have sufficient time to carry out the reading task, inefficiency in lower-level processing does not normally hinder reading comprehension, as readers are able to use strategies to compensate for processing and/or language problems (e.g. Walczyk 2000; Walczyk et al. 2007). In the present study, this hypothesis is referred to as the compensation hypothesis.

16.1.1 Theoretical Framework

The inhibition and compensation hypotheses will be explained by using two models: the Verbal Efficiency Model (VEM) (Perfetti 1988; Perfetti and Hart 2001); and the Compensatory-Encoding Model (C-EM) (e.g. Walczyk 2000; Walczyk et al. 2007).

16.1.1.1 Verbal Efficiency Model and the Inhibition Hypothesis

According to Perfetti's (1988, 1999) VEM model, subcomponents in reading are arranged in a hierarchical manner and different reading processes have ranges of processing efficiency, with lower-level processes having more potential to become automatic through extensive practice than other processes. VEM assumes that the cognitive resources (attention and WM) necessary for good reading comprehension are limited in capacity. Therefore, efficient lower-level processing skills allow cognitive resources to be used for other higher-level comprehension processes. Efficient word processing skills are central to the model and VEM maintains that the inefficient word processing skills often inhibit readers' problems with higher-level comprehension skills (e.g. building a coherent interpretation of text content, and use of reading strategies). VEM suggests that even in the adult population, for whom reading is presumably a well-practiced skill, there exists considerable variation in terms of efficiency of word processing.

16.1.1.2 Compensatory-Encoding Model and the Compensation Hypothesis

In proposing the C-EM model of reading, Walczky and his associates (Walczyk 2000; Walczyk et al. 2007) adopted basic assumptions of the VEM model but added compensatory mechanisms. Compensatory mechanisms are metacognitive in nature (Walczyk 1995), and are controlled processes, which have characteristics of slowness, serial in nature, and attention-demanding (Walczyk 2000; Wickens 1984). The establishment of the C-EM attempts to explain "the interplay between automatic and control processes" in reading (Walczyk 2000, p. 35). According to the C-EM, in fluent reading, lower-level processing tends to be carried out automatically and they make few demands on attention and WM. As a result, attention and WM can be used for higher-level comprehension processes. In situations where word processing is inefficiently or readers with a small WM, the C-EM assumes that readers are more likely turn to compensatory strategies.

One of the important predictions proposed in the C-EM model is that when there is no time pressure in reading, inefficient word processing and small WM "does not normally affect performance during reading because compensatory mechanisms operate routinely during performance" (Walczyk 1993, p. 127).

16.1.2 Empirical Evidence for Inhibition and Compensation in L1 Reading

In L1 reading, the common finding is that word processing is a good predictor of reading comprehension for beginning readers (Perfetti 1985), and word recognition among children is a major factor contributing to their later reading abilities (Adams 1999; Perfetti et al. 2005).

Whether word processing inhibits the reading comprehension of older L1 readers who are beyond the period of acquiring reading skills is more ambiguous. On the one hand, some studies showed that word processing is a good predictor for reading comprehension even among adult readers. On the other hand, other studies have found that there was only a weak link or no link between word processing and reading comprehension for older L1 readers (e.g. Walczyk 1995; Walczyk and Raska 1992).

A handful of studies with adult native English readers exist that provide some empirical evidence for the compensation hypothesis. Walczyk (1995) found that in a no time pressure condition, none of the measures of lexical access and WM (speed and accuracy) was correlated with comprehension. The results provided some evidence for the C-EM model, that under no time pressure, word processing and WM did not predict reading comprehension.

In two studies (Walczyk et al. 2001; Walczyk and Taylor 1996), Walczyk and his colleagues provided partial support for the predictions made in the C-EM. The

results of the two study showed that compensatory mechanisms (i.e. behaviours and strategies) were negatively correlated with the speed of lexical access and the speed measure of WM, suggesting that readers of inefficient lower-level processing used more frequently compensatory mechanisms. However, Walczyk and Taylor (1996) found that speed of lexical access did negatively correlate with text comprehension, meaning the faster to retrieve the meaning of words leads to better reading comprehension.

16.1.3 Empirical Evidence for Inhibition and Compensation in FL Reading

In FL reading, whether inefficient word processing inhibits reading comprehension has produced inconsistent results. On the one hand, word processing was found to positively correlate with reading comprehension in FL reading, suggesting the more proficient a FL reader process at word or sub-word level, the better he/she can achieve in comprehension (e.g. Koda 1992; Nassaji 2003; Nassaji and Geva 1999). These positive and significant correlations obtained in these studies suggest that inefficient word processing inhibits reading comprehension. For instance, Nassaji and Geva (1999) and Nassaji (2003) found that both processing beneath word level (orthographic processing) and word level processing significantly correlated with reading comprehension for adult FL readers of English speaking Farsi as their L1.

On the other hand, in other studies, word processing has been found not to influence comprehension significantly (e.g. Haynes and Carr 1990; Stevenson 2005; van Gelderen et al. 2004). For instance, Haynes and Carr (1990) found that among Chinese EFL learners, although word processing variables (i.e. word decoding variable and lexical access variable) in English reading positively correlated with reading speed, they did not correlate with levels of comprehension. Similarly, in a think-aloud study by Stevenson (2005) with Dutch adolescent EFL readers, the results suggested that word processing efficiency did not significantly correlate with levels of global reading comprehension.

16.1.4 The Present Study

L1 reading studies have showed that time pressure influences the contribution of lower-level processing to L1 reading comprehension (e.g. Walczyk 1995). Whether time pressure influences the role of lower-level processing to FL reading comprehension needs to be investigated empirically, because a handful of FL studies to date have demonstrated inconsistent results on how lower-level processing contributes to FL reading. The current study aims to investigate this issue with Chinese learners of English as a foreign language (EFL) at university level.

The study asks two research questions:

1. To what extent does lower-level processing (i.e. LA and WM) relate to reading comprehension in (a) untimed and (b) timed FL reading for Chinese EFL learners at university level?
2. To what extent does lower-level processing (i.e. LA and WM) contribute to reading comprehension in (a) untimed and (b) timed FL reading for Chinese EFL learners at university level?

16.2 Method

16.2.1 Research Design

A repeated measures design is used in which the same participant was required to read two English texts in two different reading conditions (no time pressure and time pressure conditions).

16.2.2 Setting and Participants

The study was conducted in a national university in China with 404 Chinese undergraduates (138 males and 266 females). The ages of the participants ranged from 18 to 23, with a Mean of 20.22 years old. The participants had received on average 7.5 years of English instruction: 6 years in secondary school and 1.5 year in university.

16.2.3 Instruments

16.2.3.1 Lexical Access Test

To measure word processing, the present study employed a computerized lexical access (LA) test adapted from Haynes and Carr's (1990) paper test. The LA test required learners to decide as quickly as possible whether a pair of words were synonyms or antonyms. This test was delivered using DMDX software (version 3.3.1.1), that recorded both accuracy and reaction time (RT) in milliseconds (Forster and Forster 2003).

There were 60 word pairs, half of which were synonyms and half of which were antonyms. The lexical relationship between the words was checked in an online Thesaurus (www. thesaurus.com). The reliability – Cronbach's alpha was .94, indicating a very high reliability for the LA test.

16.2.3.2 Working Memory Test

A modified computerized Operation Span Task (OSpan) (Unsworth et al. 2005), was used to measure WM. The OSpan task was delivered using DMDX software (version 3.3.1.1) (Forster and Forster 2003). There were 40 items organised into 10 sets ranging from 2 items to 6 items in one set. The test asks participants to judge simple mathematical equations at the same time to memorize isolated English words. The WM test collected three aspects of information: students' accuracy of judgment on the correctness of arithmetic equations, students' RTs on the judgment, and the number of correctly recalled English words. The three aspects were formed composite Z-scores as indicators of WM. The reliability – Cronbach's alpha was .91, suggesting the WM test was quite reliable.

16.2.3.3 Reading Comprehension Test

Reading comprehension was measured through four expository texts. In each condition, two texts were used. The four texts were adapted from *College Reading Workshop* (Malarcher 2005). Efforts were also made to maintain a similar level of text readability across the texts. Reading comprehension was measured using a multiple choice format, which is the most commonly used task for assessing reading comprehension (Brantmeier 2005; Phakiti 2008). For each text, ten multiple choice questions were constructed with four possible choices. The reading test obtained a reliability of .83.

16.2.4 Data Collection Procedure

The gathering of data was carried out in two stages. The first stage was the collection of the reading comprehension in the two reading conditions in English classes. The second stage was the collection of LA and WM data in a quiet computer laboratory.

16.2.5 Data Analysis

For research question 1, a series of bivariate Pearson product moment correlation were carried out separately for the two reading conditions. In order to answer research question 2, separate regression analyses were carried out in the two reading conditions with reading comprehension as dependent variables, and LA and WM as independent variables.

16.3 Results

16.3.1 Descriptive Statistics

Table 16.1 presents the descriptive statistics.

16.3.2 Results for Research Question 1

Table 16.2 presents the results of the correlation analyses.

Table 16.2 showed that firstly, there was no significant correlation between students' LA and reading comprehension in untimed reading condition ($r = -.09$, $p = .07$), whereas a small and negative relationship was found between LA and reading comprehension in timed reading ($r = -.22$, $p < .01$). Since RTs was used to measure LA, a negative relationship between LA and reading comprehension means that readers who were slower to access meanings of English words (who had longer RTs) tended to achieve poorly in timed FL reading; whereas readers who were faster to get access to meanings of English words (who had slower RTs) had a tendency to obtain better comprehension in timed FL reading. Secondly, WM was shown to be correlated positively with students' reading comprehension in both untimed ($r = .11$, $p < .05$) and timed reading conditions ($r = .20$, $p < .01$), and both the values of correlation were small. The magnitude of correlation between WM and comprehension in untimed reading was smaller than that between WM and comprehension in timed reading. This means that students who had larger WM

Table 16.1 Descriptive statistics for reading comprehension, LA and WM

Variables	M	SD	Min.	Max.	Highest achievable scores
Untimed reading	15.14	2.42	2	20	20
Timed reading	12.99	2.88	2	19	20
LA	1,981.06	463.59	1,026.95	3,984.58	–
WM	0.00	0.19	−0.78	0.40	–

Note: For LA test, the table reports RTs in milliseconds

Table 16.2 Results of Pearson product moment correlation analysis (reading comprehension)

Variables	LA	WM	Untimed reading comprehension	Timed reading comprehension
LA	–	**−.28****	−.09	**−.22****
WM		–	**.11***	**.20****
Untimed reading comprehension			–	**.43****
Timed reading comprehension				–

**$p < .01$; *$p < .05$ (2-tailed)

Table 16.3 Results of simple regression analysis (untimed reading comprehension)

Model	B	β	*t*	*p*	f^2
Constant	0.76	–	125.14	.00	–
WM	0.07	**.11***	**2.26**	**.02**	**.01**

B unstandardized regression coefficient, *β* standardized regression coefficient
$R^2 = .01$, $*p < .05$

Table 16.4 Results of multiple regression analysis (timed reading comprehension)

Model	B	β	*t*	*p*	f^2
Constant	0.67	–	89.05	.00	–
LA	−0.10	**−.18****	**−4.53**	**.00**	**.05**
WM	0.11	**.15****	**2.92**	**.00**	**.02**

B unstandardized regression coefficient, *β* standardized regression coefficient
$R^2 = .05$, $**p < .01$, for LA, and $R^2 = .02$, $**p < .01$, for WM

were more likely to be associated with better reading comprehension in both reading conditions. But WM had stronger association with comprehension in timed reading than in untimed reading. Additionally, LA was found to be significantly and negatively correlated with WM ($r = -.28$, $p < .01$). Lastly, reading comprehension in untimed reading was positively related to that in timed reading, and the magnitude of correlation was moderate ($r = .43$, $p < .01$). The above results between LA, WM and reading comprehension in the two reading conditions seemed to support the prediction made by the C-EM model, which maintains that when there is no time pressure in reading, inefficient word processing and small WM does not normally affect reading performance (Walczyk 1993).

16.3.3 Results for Research Question 2

Since the results of correlation analysis indicated no significant relationship between LA and reading comprehension in untimed reading, therefore, for untimed reading, only a simple regression was performed with WM as an independent variable. For timed reading, both LA and WM were used as independent variables in a multiple regression analysis. The effect size of regression analyses f^2 was also calculated and reported. The results of the regression analyses are displayed in Tables 16.3 and 16.4 separately.

The results of the simple regression analysis in Table 16.3 showed that WM was a significant predictor of English reading comprehension in untimed reading condition ($\beta = .11$, $R^2 = .01$, $p < .05$, $f^2 = .01$), accounting for only about 1 % variance, and the effect size was small. The results of the multiple regression in

Table 16.4 revealed that the variable of LA alone could explain about 5 % variance of reading comprehension in timed reading condition ($\beta = -.18$, $R^2 = .05$, $p < .01, f^2 = .05$). The variable of WM was also a significant factor for explaining the reading performance in the timed reading condition, contributing to about 2 % of variance ($\beta = .15$, $R^2 = .02$, $p < .01$, $f^2 = .02$). Both the values of the effect size attributable to the LA and WM were small, with .05 and .02 respectively. The above results suggested that the two variables, LA and WM, together could account for about 7 % variance of reading comprehension in timed-reading condition.

16.4 Discussion and Conclusion

The results of the present study seem to suggest that in untimed reading condition, LA does not inhibit reading comprehension, but in timed reading condition, the efficiency of LA inhibited reading comprehension to some extent. While there have been no studies in FL reading comparing the relationship between the efficiency of LA and reading comprehension in different reading conditions, two of the previous studies have found that the relationship between LA and reading comprehension was not significant when readers read without time imposed on them.

In a think-aloud study, Stevenson (2005) found that there was no significant relationship between the speed of English word processing and reading comprehension among 22 adolescent Dutch EFL learners. Using think-aloud protocols gave readers sufficient time for them to complete their reading tasks at hand, this situation simulated the untimed reading in the present study. However, the direct comparison of the results between the two studies needs to be made with caution. Stevenson (2005) tested the efficiency of learners' decision on whether letter strings are real or pseudo English words. This measure measurement is different to the one used in the current study, because the former measurement does not require learners to retrieve the meaning of the letter strings, while the latter measurement ask learners not only to recognize the word but also to obtain the meaning of the word from their mental lexicon.

Using the same type of LA test, Haynes and Carr (1990) found that the efficiency of LA did not significantly correlate with reading comprehension among a group of Taiwanese EFL learners. The results of the current study seem to corroborate the results of Haynes and Carr (1990)'s study among the learners with the same language background. While Haynes and Carr's study only used paper and pencil test for LA, which might not be that accurate in terms of recording RTs, the present study used computer software to capture RTs to milliseconds.

In conclusion, the results of the present study seem to support the compensation hypothesis that when reading in no time pressure condition, the lower-level processing do not seem to affect FL readers' reading comprehension.

References

Adams, M. J. (1999). Afterword: The science and politics of beginning reading practices. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading* (pp. 213–227). Oxford: Blackwell.

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *The Modern Language Journal, 89*, 37–53.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*(1), 116–124.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading*. Harlow: Longman.

Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 375–421). San Diego: Academic.

Koda, K. (1992). The effects of lower-level processing skills on FL reading performance: Implications for instruction. *The Modern Language Journal, 76*, 502–512.

Malarcher, C. (2005). *College reading workshop* (2nd ed.). CA: Compass Publishing Inc.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal, 87*, 261–276.

Nassaji, H., & Geva, E. (1999). The contribution of phonological and orthographic processing skills to adult ESL reading: Evidence from native speakers of Farsi. *Applied PsychoLinguistics, 20*, 241–267.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A. (1988). Verbal efficiency theory in reading ability. In M. Daneman, G. E. MacKinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (pp. 109–143). New York: Academic.

Perfetti, C. A. (1999). Cognitive research and the misconceptions of reading education. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading: A psychological perspective* (pp. 42–58). London: Blackwell.

Perfetti, C. A., & Hart, L. (2001). The lexical bases of comprehension skill. In D. Gorfien (Ed.), *On the consequences of meaning selection* (pp. 67–86). Washington, DC: American Psychological Association.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford: Blackwell.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237–272.

Stevenson, M. (2005). *Reading and writing in a foreign language: A comparison of conceptual and linguistic processes in Dutch and English*. Unpublished doctoral dissertation, Universitteit van Amsterdam.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505.

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed and metacognitive knowledge in first and second language reading comprehension: A componential analysis. *Journal of Educational Psychology, 96*, 19–30.

Walczyk, J. J. (1993). Are general resource notions still viable in reading research? *Journal of Educational Psychology, 85*, 127–135.

Walczyk, J. J. (1995). A test of the compensatory-encoding model. *Reading Research Quarterly, 30*, 396–408.

Walczyk, J. J. (2000). The interplay between automatic and control process in reading. *Reading Research Quarterly, 35*, 554–556.

Walczyk, J. J., & Raska, J. (1992). The relation between low-and high-level reading skills in children. *Contemporary Educational Psychology, 17*, 38–46.

Walczyk, J. J., & Taylor, R. J. (1996). How do the efficiencies of reading subcomponents relate to looking back in text? *Journal of Educational Psychology, 88*, 737–745.

Walczyk, J. J., Marsiglia, C., Bryan, K., & Naquin, P. (2001). Overcoming inefficient reading skills. *Journal of Educational Psychology, 93*, 750–757.

Walczyk, J. J., Wei, M., Zha, P., Griffith-Ross, D. A., Goubert, S. E., & Cooper, A. (2007). Development of the interplay between automatic processes and cognitive resources in reading. *Journal of Educational Psychology, 99*(4), 867–877.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). New York: Academic.

Chapter 17
Comparing Students' Citizenship Concepts with Likert-Scale

Joseph Chow

Abstract The main concern in this article has been to investigate whether Hong Kong students' citizenship concepts changed over a 10 year period. Such a methodology based on analyzing responses to the Likert scale was chosen to (1) provide precise results that make valid and reliable conclusions and (2) demonstrate that while an underlying latent trait reflecting democratic values can be identified in the two cohorts of students studied, that the latent trait did undergo subtle changes. This was reflected in the movement (or lack of it) of items along the latent trait. What is known is that Hong Kong students 10 years after the return of Hong Kong to China have strong commitments to democracy: what are less certain about is exactly how this situation developed but perhaps more importantly in the light of the results of this study, how it can be sustained.

Keywords Rasch • Likert scale • Citizenship education • Person measures • Standard error of measurement • Measurement scale • Scale score • Educational measurement • Attitudinal measurement Hong Kong • Asia

17.1 Introduction

17.1.1 Instrument and Approach

The IEA Civic Education Study (Torney-Purta et al. 2001) used 176 items to measures students' citizenship attitudes and values (Sections A to N of the Student Questionnaire). Section A consisted of 25 items related to democracy (Schulz and Sibberns 2004, pp. 246–248). Students from 28 countries were presented with 25 statements relating to democracy and were asked to determine whether the

J. Chow (✉)
The Hong Kong Institute of Education, Hong Kong, China
e-mail: chowkf@ied.edu.hk

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_17,

statement was 'good' or 'bad' for democracy. Students could choose from four response categories: 1 = very bad for democracy, 2 = somewhat bad for democracy, 3 = somewhat good for democracy, or 4 = very good for democracy. See Appendix for the specific questions.

The initial analysis of these items reported in Torney-Purta et al. (2001, pp. 71–76) did not provide sufficient evidence to report the scale properties of the items so the analysis was confined to the item level. The original student responses to these questions form part of the CivEd data base that can be publicly accessed from the website of the International Association for the Evaluation of Educational Achievement. It was these 25 democracy questions that were used in the study to be reported here. The original Hong Kong student responses provided in the CivEd database was accessed and used to survey a sample of Hong Kong students in 2009 (see below for details of these samples).

An important part of the study to be reported here, therefore, is a comparison of students' responses to the successive administrations of the items in 1999 and 2009. To ensure, legitimate comparisons could be made such, used a Rasch model containing properties that facilitated comparison on a number of key parameters (Bond and Fox 2007). Further details relating to the Rasch data analysis are provided below.

17.1.2 Sample

There were two samples used in this study. In 1999, for the IEA Civic Education Study (Torney-Purta et al. 2001), a representative sample was drawn from students between the ages of 14 and 14 and 11 months in Hong Kong. There were 4,997 students involved in the study in 1999. In 2009, as a part of a research project that aimed to investigate Hong Kong students' attitudes to citizenship in the post-handover period, over 500 Form 3 secondary school students were surveyed with the same questionnaire that was used in 1999.

For the present study, calibration samples of 500 students were randomly drawn from each of the samples of the students surveyed in 1999 and 2009. These calibration samples provided the basis for the comparisons reported here.

17.1.3 Data Analysis

As indicated above, Rasch analysis was used in this study. Its main advantage was that it transforms ordinal rating scale observations, such as responses to survey items, into interval level data the analysis of which yields more precise and accurate measurements (Bond and Fox 2007). This was an important issue for the current study. Psychometrically, Rasch analysis provided the author with scales that could be shown to be valid, reliable and comparable. Conceptually, and indeed politically, the topic of democracy is a sensitive one in Hong Kong. As academic researchers, it

is important to be sure that any claims made by this study would be based on sound evidence that Rasch analysis was capable of providing.

There is a family of analytic techniques within the broad area of Rasch measurement but this study utilized Andrich's Rating Scale Model (RSM) (Andrich 1978). The datasets were analyzed using the WINSTEPS computer program (Linacre 2009). In order to facilitate comparisons between the two student samples, the mean of the person estimates was set at zero. This became the reference point for the estimation of the item difficulties for both samples. This adoption of a zero point is different from the usual setting of Rasch analysis software but in this case it enabled the direct comparison of item difficulty estimates for the two student samples.

17.2 Results

17.2.1 Unidimensionality and Data-Model Fit

The measurement properties of the Rasch model apply only to the extent that the empirical data fit the requirements of the Rasch model. Under this model all the items which contribute to the construct being measured were included whereas items that do not contribute to the construct can be excluded (Bond and Fox 2007). Linacre (1995) suggested that there were three steps to check dimensionality: (1) addressing negative signed point bi-serial correlations since these could be a signal for obvious off-dimension behavior or a mistake in coding the data, (2) diagnosing idiosyncratic response patterns using fit statistics, and (3) identifying patterns among the standardized residuals using Rasch Principal Component Analysis (PCA) on the residuals.

In the current datasets, none of the items exhibited negative point bi-serial correlations. When the 25 items were analyzed for each sample, the fit statistics of each item were examined. In general, MnSqs within the range of 0.6–1.4 were considered as evidence of an acceptable fit to the Rasch model criteria. The results showed that all item fit statistics fell within this range, i.e. all 25 items for both samples can be said to fit the requirement for contributing to a unidimensional construct as required by the Rasch model.

PCAs of residuals were conducted on both samples and the analyses showed a distribution of items that appeared to be randomly distributed with little evidence that there were discrete sets of items clustering together to represent multiple dimensions in the data. This distribution can be seen in Fig. 17.1. It therefore seems reasonable to accept again that the 25 items formed a scale that was sufficiently unidimensional for the low-stakes analytical comparisons to be made with these data.

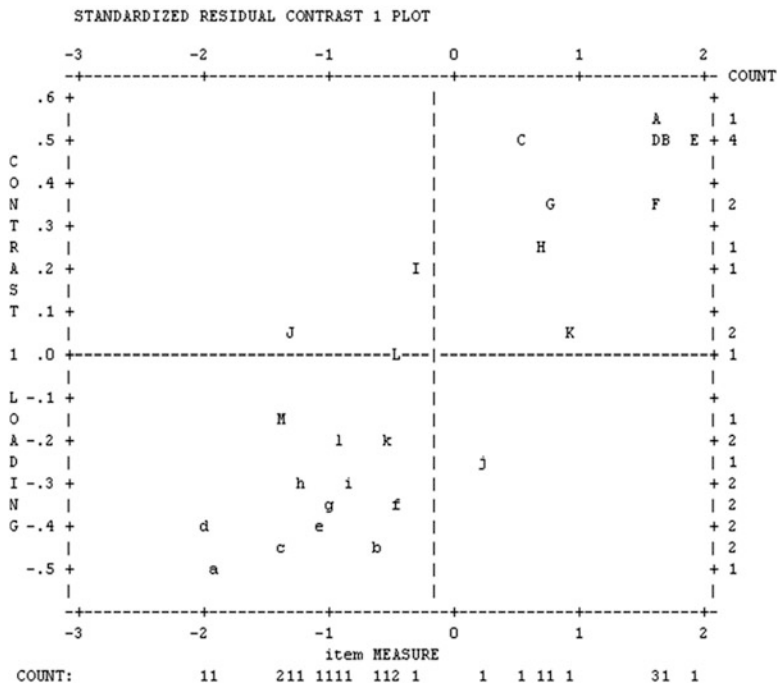


Fig. 17.1 PCA of the standardized residuals for all 25 items showing a pattern of randomness

17.2.2 Category Ordering

Since the RSM was used for data analysis, the response category orderings in the response options were examined for both datasets, revealing that the students used the categories in the manner intended by the test developers. Under the RSM, each item shared the same of category ordering. The curve of the probability for each category on the rating scale showed that there was no category disordering in the 25 items being examined (Fig. 17.2).

17.2.3 Item Statistics

Item difficulty estimates are central to Rasch analysis and they are placed on an interval scale with logits as the unit of measurement. Person measures, that are the estimates of the extent to which respondents endorse the items, are placed on the same logit (log odds units) scale so that item difficulties and person measures can be compared directly. Difficult items are those with a smaller proportion of high endorsements and consequently easier items are those with a relatively larger proportion of high endorsements.

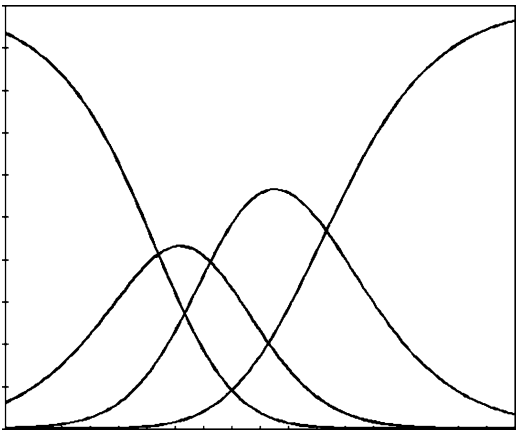


Fig. 17.2 Category probability curve of for the 25 items

For the 1999 sample, item difficulties for the 25 democracy items ranged from −1.37 logits (easiest item to endorse: A1 right to express) to 1.22 logits (most difficult item to endorse: A10 forbidden from speaking in public). For the 2009 sample, the item difficulties ranged from −1.88 logits (easiest item to endorse: A11 right to elect) to 1.80 logits (most difficult item to endorse: A12 courts and judges influenced by politicians).

The mean of the item difficulties were −0.12 and −0.13 for samples in 1999 and 2009 respectively, suggesting that overall the items in this scale matched fairly well with the student ability. This issue will be discussed further in a later section of this chapter.

17.2.4 Reliability

17.2.4.1 Item Separation Reliabilities

The item separation reliabilities, which describe how well the items were separated by the persons taking the test, were high: 0.99 for both samples. This suggests that the range of item difficulties (−1.37 to 1.22 logits for 1999 and −1.88 to 1.80 logits for 2009) was adequate in reflecting the latent trait continuum.

17.2.4.2 Person Separation Reliabilities

The person separation reliabilities, which describe the replicability of person placement across other potential items on the same latent trait, were very low: 0.57 and 0.38 for the respective sample in 1999 and 2009. This suggests that the person separation reliabilities might have been influenced by the narrow range of person estimates: 94 % of the estimates fell within the range of −1.00 to +1.00 logits. A sample with a more diverse range of abilities may not respond in the same way to these questions.

17.3 Comparisons: 1999 v. 2009

Rasch analysis produced for each of the items an item estimate with a standard error of measurement and these are shown in Table 17.1 for the 25 items for both samples. Item 2 in 1999, for example, as shown in Table 17.1, falls in the range of -0.40 ± 0.06 logits, i.e. between -0.46 and -0.34 logits. Item measures are compared across samples by comparing the magnitudes of the standardized differences between respective item measures (see Bond and King 2003).

17.3.1 Measurable Differences

In Fig. 17.3 the item estimates were plotted so that the easiest to endorse items are at the top of the graph and the most difficult below. Of the 25 democracy items, 15 items showed measurable difference. Of these, eight items (namely item

Table 17.1 Standardized differences between item estimates (with errors) for students in 1999 and students in 2009

Item	Item est. 1999	Error 1999	Item est. 2009	Error 2009	Est. 2009 – est. 1999	Combined error	Standardized difference (*t*)
A1	−1.37	0.07	−1.81	0.08	−0.44	0.11	−4.14
A2	−0.40	0.06	−0.43	0.06	−0.03	0.08	−0.35
A3	0.65	0.06	0.71	0.06	0.06	0.08	0.71
A4	−0.39	0.06	−0.55	0.06	−0.16	0.08	−1.89
A5	0.10	0.06	0.20	0.06	0.10	0.08	1.18
A6	1.10	0.06	1.47	0.07	0.37	0.09	4.01
A7	−0.81	0.07	−0.96	0.07	−0.15	0.10	−1.52
A8	0.83	0.06	1.48	0.07	0.65	0.09	7.05
A9	−0.87	0.07	−0.85	0.07	0.02	0.10	0.20
A10	1.22	0.06	1.53	0.07	0.31	0.09	3.36
A11	−1.23	0.07	−1.88	0.08	−0.65	0.11	−6.11
A12	0.98	0.06	1.80	0.07	0.82	0.09	8.89
A13	−0.90	0.07	−1.26	0.07	−0.36	0.10	−3.64
A14	−0.07	0.07	−0.47	0.07	−0.40	0.10	−4.04
A15	−0.77	0.07	−1.20	0.08	−0.43	0.11	−4.05
A16	−0.83	0.07	−1.24	0.07	−0.41	0.10	−4.14
A17	−0.34	0.07	−0.75	0.07	−0.41	0.10	−4.14
A18	−0.27	0.06	−0.41	0.06	−0.14	0.08	−1.65
A19	−0.88	0.07	−0.90	0.07	−0.02	0.10	−0.20
A20	0.25	0.06	0.52	0.06	0.27	0.08	3.18
A21	0.48	0.06	0.88	0.06	0.40	0.08	4.71
A22	0.56	0.06	0.61	0.06	0.05	0.08	0.59
A23	1.06	0.06	1.59	0.07	0.53	0.09	5.75
A24	−0.22	0.06	−0.30	0.06	−0.08	0.08	−0.94
A25	−0.77	0.07	−1.16	0.07	−0.39	0.10	−3.94
Mean	−0.12		−0.13		−0.02		

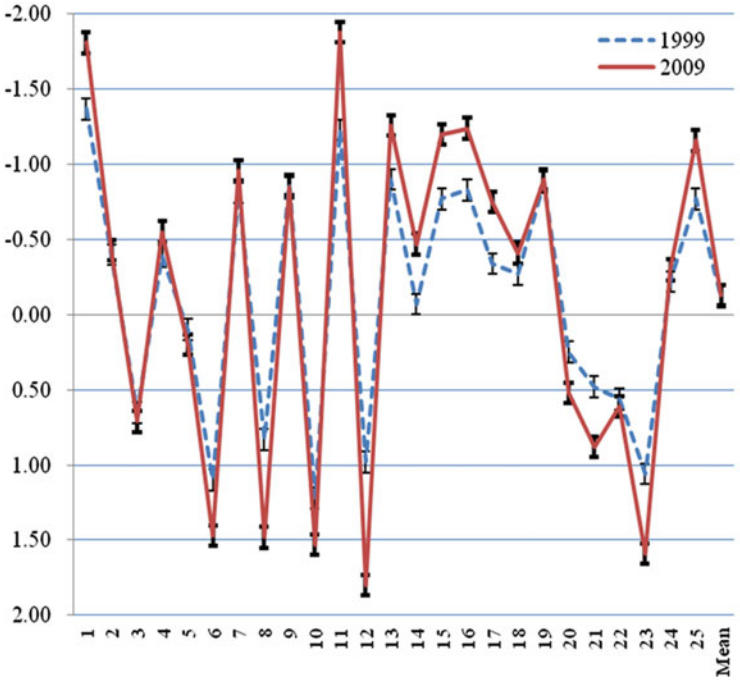


Fig. 17.3 Comparison of students' responses to democracy items showing standardized differences plotted as at t distribution (i.e. $p < .05$)

1, 11, 13, 14, 15, 16, 17 and 25) were easier to endorse (shown graphically when the points of 2009 are located *above* those of 1999). Seven items (namely item 6, 8, 10, 12, 20, 21 and 23) were more difficult to endorse for student in 2009 than for students in 1999 (shown graphically when the points of 2009 are located *below* those of 1999). The remaining ten item values were invariant (within error) from 1999 to 2009 (i.e., the error bars of the two lines overlap). Overall, there was no observable change of the scale means of item estimates across 1999 and 2009.

Table 17.1 shows the standardized differences that were computed. This approach was used by Bond and King (2003). Yet changes that are statistically significant may not always be substantively so, thus a further analysis was conducted to determine the size of the change that had taken place over the 10 year period.

Not all measureable change is necessarily substantive change, the latter meaning that there are noticeable differences in the actual size of the differences. Linacre (2008) defined a substantive differences as a difference of 0.5 logits. Newmann et al. (2001) argued that 0.6 logits corresponds to about 1-year of educational growth. Such a difference could well influence administrative decisions about students including grade retention. It seems a useful metric for effect size, therefore, when considering substantive change as opposed to measurable change. As shown

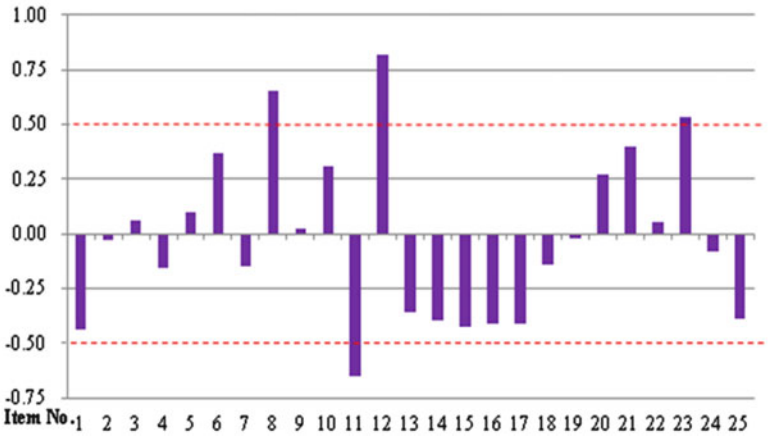


Fig. 17.4 Item differences in logit for students in 1999 and students in 2009

in Fig. 17.4 below, out of the above 15 items showing measurable differences, only four items, namely item 8, 11, 12 and 23, showed a substantive difference.

17.4 Discussion

Unlike previous analyses (Torney-Purta et al. 2001) the study reported here has used data from two samples of Hong Kong students. For both samples, the 25 democracy items can be regarded as a unidimensional scale with acceptable psychometric properties that demonstrate its construct validity. It can be assumed from the analyses provided here that the scale is an unobservable latent trait that has been tapped by the items that make up the scale. As Bond and Fox (2007, p. 311) have pointed out the observations or items themselves do not define the whole trait. Yet the items are indicative of the trait and their placement on the interval level scale serves to delineate or define as much of the unobservable latent trait as possible. The first issue to be discussed is concerned with the nature of the latent trait under discussion and the patterns of student endorsement.

The construct validity of the latent trait was established using a range of psychometric indicators. Yet it is also important to understand the underlining meaning of the latent trait. Conceptually, the common characteristic of the items used in this study is that they represent 'democratic values'. There are positive values ('good for democracy') and negative values ('bad for democracy). Thus it seems reasonable to define the 25 democracy items as the 'Democratic Values Scale' measuring an underlying latent trait. Over the 10 year period, the pattern of item location along the latent trait by the two cohorts of students was similar in some ways and different in others. First, for some items seen as 'good for democracy' the item locations remained constant for both cohorts with no measurable change (see Appendix – Items 2, 4, 7, 9, 18, 19, 24) but other similar items were

located at positions further along the relatively easy end of the scale for the 2009 cohort suggesting these items were easier for these students (see Appendix – Items 1, 11, 13, 14, 16, 17, 25). Second, some items seen as 'bad for democracy' the item locations remained constant for both cohorts (see Appendix – Items 3, 5, 22) but other some similar items were located at positions further along the relatively difficult end for the 2009 cohort suggesting these items were more difficult for these students (see Appendix – Items 6, 8, 10, 12, 20, 21, 23). The multiple directions in changes to item location together with the stability of many of the items account for the lack of overall change in the latent trait for the two cohorts of students: the mean estimate in 1999 was -0.12 ± 0.07 logits and in 2009 -0.13 ± 0.07 logits (see Table 17.1). Yet the invariance of some item locations and the change in the placement of yet other items along the latent trait does reveal something about the direction of change along the latent trait. The remainder of this section will discuss this issue.

Seven items that were relatively easy for students in 1999 and remained invariant in 2009 represented a range of basic democratic values. Most of these seem obvious: elimination of wealth disparities (Item 2), freedom of the press (Item 4), demanding political and social rights (Item 7), gender equity (Item 19 and 9) and political participation (Item 18). Yet one such item concerned trust in political leaders (Item 24) and it may not be apparent why this item should be seen as a good thing for democracy. In the original study, the item was classified as a democratic value on which there was not a great deal of consensus (Torney-Purta et al. 2001, p.74). In more traditional versions of liberal democracy trust in leaders is not always seen as a positive value and this view is often exacerbated by the confrontational nature of politics within such democracies. Yet in the Hong Kong context, and similar sociopolitical contexts in Asia, leaders tend to be more respected than they are in the West. Confucian values require that hierarchy is acknowledged and that leaders are respected. The stability of this item across time suggests that this is a value, perhaps best described as an "Asian value" that Hong Kong students have internalized. It demonstrates that at times cultural issues need to be considered when interpreting these results.

The three invariant items that were difficult for students in 1999 and remained so for students in 2009 also requires some analysis. The items were concerned with the role of government. The location of two of these items suggests that students seem to express their reservations about an interventionist role for government as even when the rationale might serve to protect minority groups (Item 22) and prevent nepotism (Item 3). Yet they do not seem to have the same reservation about government intervention in relation to business (Item 5) where they find it difficult to endorse 'no restrictions' by government. This response pattern might be suggesting that students can differentiate between different roles of government in a democracy and such a pattern might be underpinned by social and cultural considerations. For example, the role and influence of business on government in Hong Kong is an issue often discussed in the media and usually with negative connotations. In the same way, issues relating to the value of families in Confucian societies might have influenced responses to Item 3 since supporting the family is

seen as positive value perhaps even to the point of nepotism. The response to Item 22 may not so much suggest a lack of support for ethnic minorities but rather an aversion to singling out any group for special support or attention. These kind of discourses might well have influenced students' responses to these items and if this is the case then it is likely that not only are the items invariant but so too are the discourses.

These invariant items, both easy and difficult to endorse, might be seen to represent core democratic values for these two samples of Hong Kong students in this study. Thus it seems that democracy is by no means a foreign concept to these students even though neither cohort has actually experienced a democratic system either under British colonialism or Chinese sovereignty. This issue will be discussed further towards the end of the section.

It was identified four items in which there was substantive change (i.e. ± 0.5 logits) rather than just measurable or statistically significant change. One of these items reflected what was 'good' for democracy: the right to elect political leaders freely (Item 11). There was less consensus in 2009 about what was bad for democracy – forced assimilation on migrants (Item 8), political influence on the judiciary (Item 12) and business influence on the government (Item 23). The size of these changes indicates that variability in the size of the change is a factor in assessing the importance of change over time. Perhaps a better argument for the results is that the four items reported above represent substantive change of item movement along the latent trait and that the changes are in both directions. The other measurable changes may only signal emerging substantive differences between the two cohorts of students but as yet they are not of any noticeable size. One direction of change suggests more consensus about positive democratic values while the other suggests less consensus about what can harm democracy. This can be seen as an important dichotomy in the thinking of the 2009 cohort of Hong Kong students.

17.5 Conclusion

The main concern in this article has been to investigate whether Hong Kong students' attitudes to democracy and democratic values changed over a 10 year period. Such a methodology was chosen to (1) provide precise results that make valid and reliable conclusions and (2) demonstrate that while an underlying latent trait reflecting democratic values can be identified in the two cohorts of students studied, that the latent trait did undergo subtle changes. This was reflected in the movement (or lack of it) of items along the latent trait. It is shown that there are core democratic values for both samples of students. In addition it is shown that there is more support for democratic values seen to be good for democracy. On the one hand there appears to be less consensus about what can harm democracy. These changes are believed to have been brought about by the democratic discourses that characterize non-democratic Hong Kong. In other words, the lack of democracy in Hong Kong is a powerful and pervasive means that is used to both highlight and at times obscure significant democratic values. The cause maybe the media, it maybe

parents and discussions they have with their children, it may be peer interaction or it may be the availability of new technology-enhanced social networks: all of these can give students access to opportunities for learning about different aspects of democracy. What is known is that Hong Kong students 10 years after the return of Hong Kong to China have strong commitments to democracy: what are less certain about is exactly how this situation developed but perhaps more importantly in the light of the results of this study, how it can be sustained. The issue of identifying significant sources of political socialization for Hong Kong's young people remains an important research agenda for the future.

Appendix

Item no.	Item name
A1	When everyone has the right to express their opinions freely, that is
A2	When differences in income wealth between the rich poor are small, that is
A3	When political leaders in power give jobs in the government [public sector] to members of their family, that is
A4	When newspapers are free of all government [state, political] control, that is
A5	When private businesses have no restrictions from government, that is
A6	When one company owns all the newspapers, that is
A7	When people demand their political and social rights, that is
A8	When immigrants are expected to give up the language and customs of their former countries, that is
A9	When political parties have rules that support women to become political leaders, that is
A10	When people who are critical of the government are forbidden from speaking at public meetings that is
A11	When citizens have the right to elect political leaders freely, that is
A12	When courts and judges are influenced by politicians, that is
A13	When many different organisations [associations] are available [exist] for people who wish to belong to them, that is
A14	When there is a separation [segregation] between the church [institutional church] and the state [government], that is
A15	When young people have an obligation [are obliged] to participate in activities to benefit [help] the community [society], that is
A16	When a minimum income [living standard] is assured for everyone, that is
A17	When political parties have different opinions [positions] on important issues, that is
A18	When people participate in political parties in order to influence government, that is
A19	When laws that women claim are unfair to them are changed, that is
A20	When all the television stations present the same opinion about politics, that is
A21	When people refuse to obey a law which violates human rights, that is
A22	When newspapers are forbidden to publish stories that might offend ethnic groups [immigrant groups, racial groups, national groups], that is
A23	When wealthy business people have more influence on government than others, that is
A24	When government leaders are trusted without question, that is
A25	When people peacefully protest against a law they believe to be unjust, that is

References

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

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

Bond, T. G., & King, J. A. (2003). Measuring client satisfaction with public education II: Comparing schools with state benchmarks. *Journal of Applied Measurement, 4*, 258–268.

Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions, 9*(2), 422–423.

Linacre, J. M. (2008). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago: Winsteps.com.

Linacre, J. M. (2009). *WINSTEPS (version 3.69.1.7) [computer software]*. Chicago: Winsteps.com.

Newmann, F., Smith, B., Allensworth, E., & Bryk, A. (2001). *School instructional program coherence: Benefits and challenges*. Chicago: Consortium on Chicago School Research.

Schulz, W., & Sibberns, H. (Eds.). (2004). *IEA Civic Education Study technical report*. Amsterdam: IEA.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. Amsterdam: IEA.

Appendix

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7,

No. PROMS2012AU001

The Contribution of Lower-Level Processing to FL Reading Comprehension with Chinese EFL Learners

Presenter **August 8th (11:50–12:15)**

Ms Feifei Han
A PhD candidate at Faculty of Education and Social Work
The University of Sydney
e-mail: feifei.han@sydney.edu.au

Aims and Keywords

This presentation reports an ongoing research project, which investigates the two competing hypotheses: whether inefficient lexical access (LA) and small working memory (WM) inhibit text comprehension in FL reading (inhibition hypothesis) or whether readers could use strategies to compensate for processing and language problems so that text comprehension are not influenced much (compensation hypothesis).

Sample

Four hundred and two Chinese university students in their second year participated in the study.

Methods

The larger project adopted a mix-method design collecting both quantitative and qualitative data, but the focus of the presentation is only on the quantitative data. The reading text were analysed using RUMM2030 for both checking both misfitting items and persons. The data were then analyzed by doing correlation analyses between LA, WM and reading comprehension in two reading conditions: untimed reading and timed reading.

Results

Firstly, there was no significant correlation between students' LA and reading comprehension in untimed reading condition ($r = -.09, p = .07$), whereas a small and negative relationship was found between LA and reading comprehension in timed reading ($r = -.22, p < .01$). Secondly, WM showed to be correlated positively with students' reading comprehension in both untimed ($r = .11$, $p < .05$) and timed reading conditions ($r = .20$, $p < .01$), both the values of correlation were small. The magnitude of correlation between WM and comprehension in untimed reading was smaller than that between WM and comprehension in timed reading.

Conclusions

The preliminary results from the quantitative data seem to support the compensation hypothesis that when readers are allowed sufficient time, inefficient word processing and small working memory do not inhibit text comprehension in FL reading.

No. PROMS2012AU002

Generating a Learning Scale for Evidence Based Health Care: What Rasch Measurement Says Is Evidence?

Presenter **August 7th (16:25–16:50)**
Dr Ian Blackman & Ms Tracey Wachtel
School of Nursing & Midwifery
Faculty of Health Sciences
Flinders University
Adelaide, South Australia

Background
Australian health care practitioners continue to be challenged by the diverse health needs of consumers. In its mandate to prepare competent Registered nurses for the Australian nursing workforce, undergraduate nursing programs continue to articulate nursing skills development based on the concepts of evidence based practice, which in turn impacts on clinical practices.

Aims
The study seeks to identify if a scale of learning about evidence based health care practices can be generated, based conjointly on the self rated abilities of completing undergraduate nurses, to the different complexities of evidenced health care practices that are expected of contemporary nurses.

Sample
A group of Australian-based undergraduate nurses (n = 275) completing a 3 year undergraduate nursing program has been selected for the study.

Methods
A self-rated survey (using a four point Likert scale) is used by the completing undergraduate nurses to rate their self-efficacy in understanding and applying evidence based health care practices. To demonstrate this, Rasch scaling and in particular, the partial credit model has been employed. Differentiated item functioning will also be used to detect if nursing students from different subgroups score differently from each other.

Results
Outcomes indicate that indeed a robust scale of learning about evidence based health care practice can be established based on undergraduate nurses' self-rated estimates and the differing levels of complexity required for Australian evidence based health care practices.

Conclusions
The study suggests student learning about evidence based health care practices can be robustly measured using Rasch analysis. In its bid to account for as much invariance as possible, the Partial Credit Model together with Differentiated Item Functioning can give rise to reliable self- efficacy estimates for health care practitioners.

No. PROMS2012AU003

Further Implementation of ACER ConQuest User Defined Fit Statistics

Presenter **August 8th (11:00–11:25)**
Daniel Urbach
Research Fellow
Psychometrics and Methodology
Australian Council for Educational Research
19 Prospect Hill Road, Camberwell VIC 3124
t: +61 3 9277 5683
e-mail: urbach@acer.edu.au

Background

This presentation further investigates results from Adams and Wu (2004) who developed a User Defined Fit Statistic to test the fit of linear combinations of items.

Aims and Keywords

The aim of the presentation will be to showcase the advantages and uses of ACER ConQuest User defined Fit Statistics (UDFS). The advantage of using UDFS's is the ability to identify violations of local item independence and/or violations of uni-dimensionality using a-priori knowledge of item response data. In a lot of cases, individual item based fit statistics are unable to identify such violations.

Sample

Two simulated data sets both containing 1,000 cases as well as 10 and 20 items are investigated as well as two real data sets. The first comes from the Australian Graduate Medical School Admissions Test (GAMSAT) and the second, comes from a survey of Student Attitudes for a Victorian Learning Difficulties Project.

Methods

The simulated data was generated with model violations (namely local dependence and multi-dimensionality). Two real data sets were also used, one with item bundles (which was tested for item dependence violations), and one with known sub scales (which was tested for dimensionality violations). For each data set, individual item fit statistics were also compared to the UDFS's. This was done in ACER ConQuest.

Results

The simulations showed the UDFS's ability to identify violations of item response model assumptions, when local item dependence and multi-dimensionality were introduced in the generated simulation data. Two real datasets were also analysed, and such violations were again identified.

Conclusions

When a priori knowledge of item combinations of instruments are known, which is often the case in commercial assessments which contain item bundles or sub-scales, UDFS's may be extremely useful in identifying violations of local item independence and uni-dimensionality in practice.

No. PROMS2012AU004

Investigating the Measurement Invariance in International Assessment Programs Using the Rasch Simple Logistic Model

Presenter
Rassoul Sadeghi, PhD
Sofia Kesidou, PhD
Educational Assessment Australia (EAA)

August 6th (13:55–14:20)

Participation in international assessment programs that are designed to monitor student progress has increased in recent years. Such assessment programs provide important information on the achievement of students in a specific country relative to students in other countries, and identify factors that have contributed to this achievement. Since curriculum and language of instruction varies across different participating countries, it is often essential for the original tests to be adjusted for curriculum differences and be translated into different languages. However, it is widely recognised that curriculum differences and differences between source and target languages may have a significant impact on the equivalence of test items. As a result, determining the cross-lingual and cross-country equivalence of the international test items is fundamental. This study is designed to investigate the impact of language and curriculum Differentional Item Functioning (DIF) on the construction of an International Science Scale (ISS). A total of 12,762 students (Grade 4–Grade 10) took a Science Test in 2011. Subjects were classified into three groups: (1) English native speakers who took the English version, (2) Chinese native speakers who took the Chinese version and (3) Chinese native speakers who took the English Version. The Rasch Simple logistic model was applied to analyse the data using RUMM 2030 software. RUMM 2030 uses the analysis of variance of the residuals to identify those items with language DIF. Once DIF was detected, the 'top-down purification' approach (Tennant and Pallant 2007) was adopted to deal with items displaying DIF. The results of this study show that DIF due to curriculum and language can change the relative position of examinees on the latent variable. This study has important implications for international assessment programs, particularly when data obtained from different countries are to be pooled to construct a common scale.

No. PROMS2012AU005

Extended Mantel-Haenszel Procedure in DIF Detection

Presenter **August 8th (9:50–10:15)**
Dr. Xiaoxun Sun
Psychometrics and Methodology Research Program
Australian Council for Educational Research
19 Prospect Hill Road, Camberwell, VIC, 3124
t: +03 9277 5419
e-mail: xiaoxun.sun@acer.edu.au

Background
Many methods have been developed for studying Differential Item Functioning (DIF). An approach developed by Holland and Thayer (1988) adapted the Mantel-Haenszel (MH) statistic (Mantel and Haenszel 1959). Xiaoxun Sun has implemented MH procedure and its extension in an experimental version of ACER ConQuest which can handle DIF analysis for multiple focal groups and polytomous items.

Aims
This presentation compares the extended MH procedure in DIF detection to the standardized item difficulty difference method based on the Rasch estimates for both dichotomous and polytomous items. The use of MH method in DIF analysis with multiple focal groups will be illustrated.

Keywords DIF • MH statistics • ACER ConQuest

Sample
Grades 3–10 data from International School Assessment (ISA) are used in the study. The assessment data were from 25,000 students.

Methods
The DIF analysis results based on the standardised difference method has been compared with results using extended MH procedure. In using the MH procedure, stratification based on ability is studied in terms of appropriate number of strata. ACER ConQuest has been used for the analyses.

Results
The Rasch DIF method has a higher DIF detection rate than the MH procedure. The appropriate number of strata used in MH method is found to depend on the test length and ability distribution.

Conclusions
The MH results showed a more conservative DIF detection rates compared with the standardized difficulty difference method except for very easy or very difficult items. The MH method can be used to detect non-uniform DIF. The extended MH procedure can better handle multiple focal groups and polytomous items.

No. PROMS2012AU006

The Attitudes of School Leaders to the Relationship Between School Registration and School Improvement

Presenter **August 8th (10:15–10:40)**
Harm (Pete) Witten
Doctor of Philosophy Candidate
Faculty of Education and Arts at Edith Cowan University
Principal Supervisor: Professor Russell F. Waugh
Associate Supervisor: A/Professor Jan Gray

In 2004, the Government of Western Australia introduced a mandatory inspection-type registration process for all Non-Government Schools. Part of the aim of this registration process was to help schools improve twelve educational and administrative aspects. These were: (1) School Governance, (2) School Financial Viability, (3) Enrolments & Attendance, (4) Number of Students, (5) Instructional Time, (6) School Staff, (7) School Infrastructure, (8) School Curriculum, (9) Student Learning Outcomes, (10) Care for Students, (11) Disputes and Complaints, (12) Legal Compliance. A questionnaire based on these twelve aspects was designed with five items per aspect (60 items total), conceptually ordered from easy to hard, and given to 110 administrators. It was completed by 65 administrators for a useable, response rate of 59 %. The data were analysed with the Rasch model computer program RUMM2030 which accommodated the small numbers by estimating parameters even when some response cell frequencies are zero or low. It does this by re-parameterising the thresholds into principal components (not the factor analysis kind), but components that make up the structure of the threshold parameters where there are data. The frequencies are not used directly, but rather functions of the frequencies are used as the sufficient statistics for these parameters and the thresholds are recovered from these. A unidimensional, linear scale, School Administrators' Beliefs That Actual School Improvements Were Due to Formal School Registration, was created with 48 items. The Person Separation Index of 0.86 was highly satisfactory. The item-trait interaction was 83.76, $df = 96$ with $p = 0.81$ supporting the creation of a unidimensional scale. The results showed that there was a group of items that administrators said were relatively easy to say that actual school improvements were due to formal registration and another group that administrators said were very hard to say that actual school improvements were due to formal registration. This study produced a new Rasch measurement for a key aspect of school improvement. It provides new insight into the policy and practice of school registration.

No. PROMS2012AU007

The Impact of Unobserved Extreme Categories on Item and Person Estimates: A Simulation Study

Presenter **August 8th (11:00–11:25)**
Dr. Edward Feng Li
Educational Assessment Australia
UNSW Global
e-mail: edwardfli@hotmail.com.

Background

For any polytomous items, it sometimes occurs that an extreme category, which is logically possible, is not observed in a particular sample. For example, in education when the performance tasks by the students from different year levels are judged by the same set of criteria, it is likely that none of the lower year level students would achieve the highest marks in some criteria. In health, it may happen when a group of generally healthy participants are measured by an instrument designed to detect some particular symptoms.

Aims and Keywords

This paper uses a simulation study to investigate the impact of unobserved extreme categories on item and person estimates.

Keywords Polytomous items • Unobserved extreme categories • Item estimates • Person estimates

Sample

The sample size for this simulation study is 1,000, with a mean of 0 and a standard deviation of 1, N(0,1).

Methods

Based on the polytomous Rasch model, the Partial Credit Model (Master 1982), data were simulated for 1,000 persons and 10 polytomous items with five categories under two scenarios, one with unobserved extreme high categories and the other with unobserved extreme low categories. The generated data sets were analysed with the RUMM2030 software (Andrich, Lyne, Sheridan and Luo 2009).

Results

The results show that unobserved extreme high categories in the data tend to lead to underestimated person measures and unobserved extreme low categories in the data tend to lead to overestimated person measures. The results suggest that collapsing unobserved extreme high categories, even when no disordered thresholds are present, improves person and item estimates, compared with leaving them

unchanged. However, collapsing unobserved extreme low categories only improves item and person measures when a large proportion of items have unobserved extreme low categories.

Implication
These results have implications for designing and measuring performance tasks that need to be carried out across a wide spectrum of ability groups.

No. PROMS2012AU008

A Comparison of Ratings Between American and Australian Nurses Using Differential Item Functioning (DIF)

Presenter
Dr Patricia Nicholson
Melbourne School of Health Sciences, Faculty of Medicine
Dentistry & Health Sciences
The University of Melbourne
e-mail: pfnich@unimelb.edu.au

Background

Following validation of a Performance Based Scoring Rubric using Item Response Modeling the analysis showed that the Analytical Observation Form produced acceptable reliability estimates (0.94) and the fit statistics for most of the items were acceptable. Using a video-clip as a fixed prompt, the factors likely to influence the accuracy of the performance ratings of nurse educators were explored using the validated instrument.

Aims and Keywords

The aim of the study was to determine to what extent the ratings of nurse educators were influenced when using an Analytical and Holistic rubrics to judge the performance of an instrument nurse observed in a video-clip.

Keywords Differential item functioning • Rasch Model • Competency-based performance

Sample

The sample included nurse educators who were involved in providing clinical support to nurses working in the operating theatre (Australia = 186; America = 127).

Methods

A group of nurse educators were required to observe the performance of an instrument nurse captured in a video-clip using the rubric which was developed using the Australian College of Operating Room Nurse and Association of peri-Operative Registered Nurses unit of competency and standards of practice. An independent-samples *t*-test was conducted to compare the mean scores of American and Australian nurse educators with a significant difference ($p < 0.001$)in the mean scores found (26.46 and 22.74 respectively). In order to explore whether errors or bias were impacting on the quality of the data, the *t*-statistic of the paired item difficulty estimates (i.e. American and Australian item difficulty estimates) for each item on the Analytical Observation Form was explored using ConQuest (Wu et al. 2007). A direct comparison was then made to determine the differences in each subgroup.

Results

Although there were eight items with *t*-statistics smaller than −2 and larger than 2, only two items exhibited significant differences with *t*-values of −3.96 and −7.238 respectively.

Conclusion

The study had several limitations which included the use of a video-recording of practice. Although DIF was detected between the Australian and American nurses, further research is required to confirm the findings from this study and allow for the generalization of the findings in clinical practice. The use of international competency standards also have wide reaching implications for further research in examining nursing practice difference in various countries.

No. PROMS2012CN001

A Research on the Effectiveness of DynEd Computer-Assisted English Language Learning: Taking Ningbo Polytechnic as an Example

Presenter **August 8th (11:25–11:50)**
Jingru Huang; Baixiang Wu
Ningbo Polytechnic
Ningbo, Zhejiang, China

The Integration of Multimedia-aided English language teaching and learning is one of the trends of abroad and domestic research, and DynEd CALL (computer-assisted English Language learning) within the framework of integration is a new challenge. This paper explores the DynEd CALL model based on public English teaching reform of Ningbo Polytechnic that has been done for 3 years, which aims to find the effective model of language learning and teaching. The purpose of this study is to investigate the effectiveness of DynEd CALL model. The data were collected through DynEd Records Manager, on over 2000 non-English majors from 7 institutes of Ningbo Polytechnic, implemented for 1 year. The data includes each student's studying time, days studied, two placement tests, two speaking tests (both were done at the beginning of learning DynEd and at the ending of learning DynEd), study score, mastery tests score of each unit and detailed studying information about the courses of *New Dynamic English* and *First English*, such as functional buttons usage and speech recognition, etc. The data also contains a questionnaire to each student followed by 1 year's DynEd study, 30 questions using Likert 5-point scale. Pre-and post-comparative analyses of two placements tests and speaking tests show that the differences are nearly 0.5 and 0.4 respectively, which means the award-winning, multimedia content keeps students on task and engaged. And the questionnaire indicates students have adopted the concept that language is a skill, not knowledge. DynEd CALL can significantly promote students autonomy and students' listening and speaking skills, but deficiency on reading and writing skills, which are the perspectives we should enhance in the following teaching.

Keywords DynEd computer-assisted • Language learning • Effectiveness • Data base

No. PROMS2012CN002

Validation of a Large-Scale Reading-to-Write Task: Evidence from Multi-faceted Rasch Model Analysis

Presenter
Zhang Xinling, PhD
School of Foreign Languages
Shanghai University
Shanghai, China

August 6th (15:55–16:20)

The new decade witnesses the advent of integrative writing tasks in large-scale tests home and abroad, with accompanying validity evidence (Asencion 2004; Messer 1997; Watanabe 2001; Xu Hao and Gao Caifeng 2007; Zhang Xinling and Zeng Yongqiang 2009). However, research adopting Rasch model to detect test score variances are fairly inadequate. The present study employed MFRM to collect validity evidence for a large-scale reading-to-write test task. Analyses of 190 subjects' writings with FACETS and the formula of $\log(P_{njmik}/P_{njmi}(k-1)) = B_n - C_j - E_m - D_i - F_{ik}$ yielded the following findings: (1) The task can distinguish among candidates of different reading-to-write abilities, and score variance is largely attributed to the construct; (2) The task is difficult for the targeted candidates; (3) bias analysis indicated that differentiating rating behavior of individual raters necessitates rating rubrics improvement and further rater training.

Keywords Reading-to-write tasks • Multi-faceted Rasch model • Validation

No. PROMS2012CN003

An Analysis of Item and Person Fit to Rasch Model

Presenter **August 8th (9:25–9:50)**
Guoxiong He[1], Huifeng Mu[2]
Faculty of Foreign Studies, University of Jiaxing
Jiaxing, Zhejiang Province
P. R. China
e-mail: 1. Hgx505@mail.zjxu.edu.cn
2. mhf775@yahoo.com.cn

Aims and Keywords
The study is to analyze item and person fit to the Rasch model, aiming to improve test quality and facilitate item banking.

Keywords Rasch • Fit analysis • Item difficulty

Sample
Five hundred first-year non-English major students from Jiaxing University, 18 years of age and up, male and female, are randomly sampled in the study. The scores were obtained from a reading subtest, which consists of 30 multiple-choice items.

Methods
The scores from 500 candidates on 30 items were arranged in successive order of difficulty and ability. GITEST III + SYSTEM was used to calculate item difficulty and candidate ability. For each item and each test taker, Rasch model fit statistics were calculated, including Wright's squared standardized residuals (W) and associated t values. In the meantime, we split up the candidates into six groups on the basis of average ability. Bock's chi-square (BCHI) was used in order to check whether the observed data can fit the IRT one parameter. We also take the natural logarithm of the likelihood function at the max. As this is dependent on the level of theta, so we need it standardized, which is Drasgow's standardized appropriateness index (l_z).

Results
The results given by GITEST III + SYSTEM show that 6 items and 40 candidates were identified as misfitting by the conventional criterion of fit surpassing 2.00. But BCHI and l_z produce the outcomes with different number of misfitting items and candidates, suggesting that several procedures are needed to check the degree of misfit.

Conclusion
No single fit statistic or statistical test can fully interpret item and person fit to the model because different methods produce different results. Some fit statistics are very sensitive to changes in sample size. We need several procedures for checking the degree of misfit, the unlikelihood of success or failure in responding to a test item.

No. PROMS2012CN004

The Washback Effect of TEM-4 on Teaching of English Majors at Sport Universities and Institutes in China

Presenter
Han Bing, Liu Lirui
Foreign Languages Department
Beijing Sport University
Beijing, China

August 7th (16:00–16:25)

The primary goal of objective language testing is to make an accurate and fair measurement of the language users. As a monitoring and evaluation mechanism, it maintains a positively interactive relationship with language teaching. In China, objective language testing has been widely used in some nation-wide tests, such as the Test for English Major (TEM-4), due to its quantitative design, convenient scoring and favoring fairness. Based on Alderson and Wall's 15 washback hypotheses (1993) and empirical washback studies, the paper set out to investigate how effectively and efficiently TEM-4 can measure students' competence in language learning and how it should be used appropriately in classroom teaching. Data were collected from over 525 teachers and students majoring in English of 10 sport universities and institutes in China by means of a questionnaire survey and in-depth interviews. Findings from this study indicated TEM-4 produced more positive washback effects than negative ones in that students improved their learning strategies and language proficiency; TEM-4 offered teachers feedback and helped them foster students' comprehensive English abilities by using the test as a motivation tool. However, some discrepancies further supported that the washback effect was quite context-oriented and complicated. In sum, the paper achieved a breakthrough in carrying out a mixed (quantitative and qualitative) method from the perspective of teachers and students to investigate washback in the less explored area of Chinese sport universities and institutes, hoping the results could make some contribution to the improvement of English teaching in the sport universities and institutes in China.

Keywords Objective language testing • Washback effect • English teaching at sport universities and institutes • TEM-4

No. PROMS2012CN005

Implementing Formative Assessment in the Translation Course for English Majors

Presenter
Siqi Lv
English Department
Beijing Sport University
Beijing, China
e-mail: bsu_kenny@foxmail.com

August 7th (13:50–14:15)

Educational assessment as well as studies on it has been playing a pivotal role in language teaching since the concept of "educational assessment" was put forward by R.W. Tyler in the 1930s. As more and more emphasis is put on the development of students' ability and the process of teaching, formative assessment, which differs from the traditional result-oriented summative assessment, is being widely applied.

Although the benefits brought by and importance of formative assessment have been confirmed, most of the researches are carried out only in college English classes, with little practice in English majors. This thesis intends to implement formative assessment in English majors' translation course, aiming at finding out its effects on students' translation ability as well as learning capabilities, and further providing suggestions on future studies in this area. Forty junior English majors from Beijing Sport University have been chosen for the research, during which self-assessment, peer-assessment and teacher assessment are used to implement formative assessment. Assessing tools including two achievement tests, journals, classroom observation, questionnaires and in-depth interviews have been used to collect both the quantitative and qualitative data. The research finds out that students' role has changed from a listener to a participant, while the teacher from an authority to a guide. Students' translation ability has been improved after the research and they have made progress in other learning capabilities like self-reflection, critical thinking and learning autonomy, which are crucial for future studies in other areas.

Keywords Formative assessment • Translation course • Translation ability • Learning capability

No. PROMS2012CN006

Measuring the Quality of Teacher-Child Interaction in China's Preschools: A Rasch Measurement Approach

Presenter **August 6th (13:55–14:20)**

Xiaoting Huang, Yingquan Song and Loyalka Prashant
China Institute for Educational Finance Research
Peking University
R404, College of Education
Peking University, Beijing, China
phone: 010–62758588
e-mail: xthuang@ciefr.pkue.du.cn

In recent years, Chinese government has drastically increased the national financial allocation on early childhood education (ECE), as a result of a growing consensus regarding the importance of ECE. Despite the rapid expansion, little attention has been brought on the quality of this sector. Previous research found that quality in ECE has many dimensions, including the "structural" and "process" aspects. Most of the studies in China focus on the structural aspect. Little has been said about the process quality. In this study, we constructed an instrument to measure teacher-child interaction-based process quality via Wilson's four-building-blocks approach. Specifically, we designed an 11-item observation protocol. Trained raters observed each class for about 30 min before they scored the teachers. Eight hundred and eighteen teachers were observed in July 2011. The data were calibrated using Rasch-based partial credit model. The initial result showed that the overall reliability was only moderately high at 0.7. The scores were positively correlated with principal ratings, as well as teacher certification and professional training experience. The findings suggest that process quality can be scientifically measured using carefully designed instruments. The instrument needs be improved in the future by adding more items and refining descriptors, etc. Finally, more elements of the process quality should be incorporated.

No.ROMS2012CN009

Multi-facet Rasch Model Applied to Investigating Rater's Effects in PETS-SET

Presenter **August 6th (14:20–14:45)**

Laura Yang Hong*, Mingzhu Miao and Yan Zhao
Faculty of Foreign Studies, University of Jiaxing
56 Yuexiu South Rd, Jiaxing, Zhejiang Province, China
e-mail*: yanghong_158@yahoo.com.cn

Background

Public English Test System (PETS), one of the large-scale tests in China, is a proficiency test. The purpose of its Spoken English Test (SET) is to measure the ability of the test takers to communicate orally in English. It has a higher face validity. **However**, the raters' subjective judgments and other problems of the raters' instability and differences can affect the scoring, resulting in inconsistency of test scores, i.e. low reliability. Therefore, this study is to provide some implications for rater training so as to improve the rater reliability.

Aims and Keywords

Based on the modern Item Response Theory (IRT), the Multi-facet Rasch Model was used to investigate four types of rater effects: severity, rater instability, halo effects and extremism/central tendency to find out the rater reliability.

Keywords Rater effects • PETS-SET • Multi-Facet Rasch Model • Reliability

Sample

There were over 400 examination candidates took PETS-SET Band 2 in Jiaxing City in 2010. Eight pairs of raters and their ratings were chosen randomly, with each pair rating 30 candidates. The assessor adopted a 5-point analytic rating scale which consisted of three separate domains: Grammar and Vocabulary, Pronunciation, and interactive Communication. The interlocutor adopted a 5-point rating scale for Global Achievement.

Methods

The ratings were analyzed by using a Multi-facet Rasch Model (MFRM) conducted by the software package FACETS. The severity levels and the potential halo effects of the raters were investigated by using the primary analysis of MFRM and then the rater's stability.

Results

Raters showed significant differences in their severity level; Raters behaved consistently in rating but some raters showed bias towards certain domains; and only 2 raters showed significant halo effects, i.e. they tended to give similar scores.

Conclusions

The results show statistically significant differences among all facets including severity, rater instability, halo effects and extremism/central tendency. MFRM is an effective means for measuring the rater effects.

No. PROMS2012CN010

Language Aptitude Components and Different Levels of Language Proficiency Among Chinese English Majors

Presenter **August 7th (13:25–13:50)**
Lanrong Li
Beijing Normal University
e-mail: Jessicallr@mail.bnu.edu.cn

Background
Foreign language aptitude is recognized as one of the most important individual variables in second language acquisition. Previous studies suggest that language aptitude is componential and different aptitude components may play different roles at different levels of second language proficiency. However, relatively little research to data has been conducted to test this hypothesis.

Aims
This study aims to explore the relationship between language aptitude components and different levels of English proficiency among Chinese English majors.

Sample
Sixty-two second-year English majors (9 males and 53 females, aged between 18 and 23) from a university in Beijing participated in the study.

Methods
An aptitude test composed of subtests of Pimsleur Language Aptitude Battery and subtests designed by the author was administered to the participants. The students' TEM-4 and TEM-8 scores were used as measures of their English proficiency. Correlation analysis, multiple regression analysis and ANOVA were conducted.

Results
Results showed that different aptitude components had different relationship with the two language proficiency tests. Regression analysis showed that two aptitude components (Sound Discrimination and Memory for Text) were significant predictors of the students' TEM-4 scores while three aptitude components (Language Analysis, Sound Discrimination and Memory for Text) were significant predictors of the students' TEM-8 scores. Further analysis showed that students with higher and lower TEM-4 and TEM-8 scores also differed significantly in different language aptitude components.

Conclusion
The results lend support to the hypothesis that different language aptitude components may play different roles in second language acquisition when the learner is at different levels of proficiency, which implied that instructional methods could be tailored to suit learners' language aptitude profiles according to their proficiency levels.

No. PROMS2012CN018

A Pilot Study Based on Rasch into the Appropriateness of the TOEIC Bridge Test for Chinese Students: Status Quo and Prospect

Presenter **August 7th (16:00–16:25)**

Quan Zhang
Member, Supervisory Committee of ELT in Vocational Higher Education, Ministry of Education, P.R.China and Faculty of Foreign Studies, University of Jiaxing, P.R. China
e-mail: qzhang141@yahoo.cn

Mingzhu Miao & Chunyan Zhu
Faculty of Foreign Studies, University of Jiaxing, P.R. China

Eng Han Tan
Head of ETS Beijing Rep Office
e-mail: etan@etsglobal.org

Background

Nowadays, the number of Chinese vocational students enrolled each year is increasing. Up to the present (2009), the total number nationwide amounts to almost 3.3 millions, of whom approximately 1.5 millions are English majors. However, over the past decades, there was no a well-accepted English language test for such a big number of students. TOEIC *Bridge*™ test (http://www.ets.org/toeicbridge) is being paid attention to by both Chinese government and university educators. It is the only English test officially introduced into China by Ministry of Labor and Social Security (MOLSS) under the Chinese government. To validate the test scores from the TOEIC Bridge as a measurement of English proficiency for and to eventually replace local tests of various kinds currently administered to Chinese vocational students annually of over 3.3 millions, the Supervisory Committee of ELT in Vocational Higher Education, Ministry of Education, P.R. China decided with much care to initialize the pilot study. The present study is part of this project.

Aims and Keywords

The purpose of conducting such a study is in an attempt to achieve at least four goals as follow in terms of (1) the appropriateness of TOEIC Bridge Test to Chinese vocational students and a research report to department of higher education for reference; (2) possible separation of teaching from testing; (3) improvement of test quality and (4) recognition of test certificate.

Sample
Students of Jiaxing University and other universities wihtin Zhejiang Province.

Method
Comparison of TOEIC Bridge Test and other relevant English test in China using Gitest, Rasch-based software.

Results and Conclusions
The results thus obtained show that TOEIC Bridge test fits Chinese students better and the score interpretation better indicates Chinese students' communicative competence in the real social and campus life. Evidences by comparison collected via questionnaires also indicate the possible reasons why the local parallel test is not very well accepted by educators and employers in China.

No. PROMS2012CN019

Investigating the Consequences of the Application of Formative Evaluation Reading-Writing Model

Presenter **August 7th (15:35–16:00)**

Laura Yang Hong*, Yan Zhao and Qiu Yulei
Faculty of Foreign Studies, University of Jiaxing
56 Yuexiu South Rd, Jiaxing, Zhejiang Province, China
e-mail*: yanghong_158@yahoo.com.cn

Background

Nowadays, we can see reading accounts for a large proportion in some large-scale English proficiency tests at home and abroad, such as CET-4, CET-6, TEM-4, GRE, IELTS, and TOEFL. Many studies show that a considerable number of Chinese students can get high marks but they have never developed true reading skill and one of the major reasons is that they lack adequate effective reading. How to combine reading and writing effectively in teaching has become an important research field in Second Language Acquisition (SLA) in China.

Aims and Keywords

The purpose of conducting such a study is in an attempt to answer the following questions: (1) Can formative evaluation be helpful in anchoring learners' identity (attitudes and learning needs in reading)? (2) Can formative evaluation be helpful in students' learning management (i.e. learning strategies)? (3) Can formative evaluation be helpful in promoting students' writing and reading effectively?

Keywords Formative evaluation • Reading through writing model (RTWM) • Consequence

Sample

There were 35 English Majors from Jiaxing University who had the course Extensive Reading. They were assigned an integrative task — choose to read some English newspapers and magazines two or three times a week and keep a notebook, and choose two original novels a semester and write book reports on storyline, theme, writing style, favorite hero(s)/heroines, comments, what they get, and so on.

Methods

This study adopts a comparative way, in which two freshman English major classes with similar motivation, strategy and proficiency, sex distribution and taught by the same teacher will be engaged. Class A, where RTWM is practiced, is the experimental group and Class B with the conventional assessment is the control group. The data elicited from interview, observation and students' notebook are used for qualitative analysis and the data from questionnaires and reading and writing tests are used for quantitative analysis by using the software SPSS 15.0.

Results and Conclusions

The results thus obtained show that the application of formative evaluation to RTWM is of great significance in anchoring learners' identity and improving students' learning management in that it is the students who can tell the teacher what they have learned and what they are learning. Evidences by comparison collected via questionnaires and tests also indicate that formative evaluation be helpful in promoting students' writing and reading effectively.

No. PROMS2012HK001

A Rasch Analysis on the Development and Validation of Mathematics Test for Use by Primary Five Student in Hong Kong

Presenter **August 7th (15:10–15:35)**

Jingjing Yao[1], and Magdalena Mo Ching Mok[1,2]
[1]Assessment Research Centre, Hong Kong Institute of Education
e-mail: jingjing@ied.edu.hk
[2]Psychological Studies, Hong Kong Institute of Education
e-mail: mmcmok@ied.edu.hk

Background

This study was part of a large-scale quasi-experiment on the causal relations among feedback, self-directed learning, and mathematics achievement of primary students in Hong Kong. The larger study was sponsored by a General Research Grant awarded by Research Grants Council of UGC, Hong Kong (Project Number 844011).

Aims and Keywords

The study aimed to develop and validate, using a Rasch approach, of an instrument for measuring mathematics achievement of Primary five students.

Keywords Rasch model • Validity • Mathematics achievement scale • Primary student

Sample

The sample comprised 1,368 Primary 5 students from 16 Hong Kong schools who participated voluntarily.

Methods

A pre-test/post-test design was used for the larger study. Data for this study were collected at the pre-test in form of a mathematics test completed during class-time. A 35-item mathematics test was developed after careful analysis of the local curriculum, and in consultation with teachers on the suitability of the test for students by the end of Semester One of Primary 5. Rasch analysis with the Winsteps software (version 3.70) was then used to validate the instrument.

Results

The results showed that: (1) item-fit was between 0.5 and 1.5 for all items except one; (2) Eigen value of the first contrast of the Rasch PCA of Residuals was below 2.0; (3) Rasch person and item reliabilities were 0.80 and 1.00 respectively; (4) there was good alignment between item difficulty and student ability; (5) no gender-DIF was found in the items.

Conclusions

Measurement validity is one of the most crucial factors for the overall validity of any research study. This study developed a valid and reliable instrument for measuring mathematics achievement of Primary 5 students.

No. PROMS2012 HK002

The Mixture Facets Model for Differential Rater Functioning

Presenter **August 6th (16:20–16:45)**
Jin Kuan-Yu, Wang Wen-Chung
The Hong Kong Institute of Education
Hong Kong

Background
Essay items have been widely used in educational tests. Often, raters need to be recruited to mark essay items and raters may have very different degrees of severity. Furthermore, a rater may show different degrees of severity for different groups of examinees, which is referred to as differential rater functioning (DRF). In most DRF studies, the group memberships of examinees are known, for example, gender or ethnicity. However, DRF may occur when the group memberships of examinees are unknown.

Aims and keywords
This study aims to develop a new mixture facets model to assess DRF when the group membership is unknown.

Sample
Simulated datasets were generated with two latent classes by manipulating six conditions: (a) sample size; (b) group mean difference; (c) number of items; (d) number of raters; (e) magnitude of DRF; and (f) tendency of DRF.

Methods
The generating model was fit to the simulated datasets. The dependence variables were the bias and root mean square error of parameter estimates, and the accuracy of identification of latent classes. Twenty replications were made under each condition. A Bayesian method was used to estimate the parameters by using the WinBUGS freeware.

Results
The parameter estimation and the classification of latent classes were more accurate when the dataset was larger (i.e., larger sample sizes, longer tests, and more raters), the group mean difference existed, the differences in the rater parameters between latent classes were larger, and the pattern of DRF was balanced between groups.

Conclusions
Fitting the proposed mixture facets model is useful to explore the inconsistency of rater severity with respect to different latent classes.

No. PROMS2012HK003

Multiple Regression of Civic Knowledge Using Plausible Values

Presenter **August 8th (11:50–12:15)**
Kuang Xiaoxue
PhD student in Hong Kong Institute of Education
e-mail: Kuangxiaoxue2006@126.com

Professor KENNEDY, Kerry John
Chair Professor of Curriculum Studies, Associate Vice President (Quality Assurance), Dean of Faculty of Education Studies, Co-Director of Centre for Governance and Citizenship
e-mail: kerryk@ied.edu.hk

Aims and Keywords

The purpose of our study is to find out whether students' views on values and attitudes underlying citizenship issue can be used as the variable to predict their civic knowledge.

Keywords ICCS 2009 • Plausible value • Multiple regressions

Sample

The data used for this study were retrieved from the ICCS database. The samples used in the International Study were also sued for the Asian Regional Module (ARM). Sample sizes for each participating society are shown in Table 1. The sample was drawn from Grade 8 students between the ages of thirteen and fourteen. The average age of the sample was 14.4 years with student ages ranging from 14.2 in Chinese Taipei to 14.7 in Korea.

Table 1 Sample size for the five Asian Societies participating in the ARM

	Chinese Taipei	Hong Kong SAR	Indonesia	Korea, Republic	Thailand
Sample	5,152	2,739	5,048	5,252	5,263
Valid	5,152	2,737	5,048	5,252	5,263

Methods

Multiple regression analysis has been used to interpret the relationship among the Asian Regional Module (ARM) and civic knowledge. The predicted variables are the three latent construct of the ARM—students' perceptions of government and law in Asia, students' perceptions of identity, citizenship, and culture in Asia students' perceptions of public service for student's civic knowledge. In order to do the analysis, the plausible value of the three predicted variables of five Asian countries are computed using ACER Conquest software (von Davier, Gonzalez, and Mislevy 2009). Since there are five plausible values, the multiple regression are

repeated five item. The final result is the average of the five regression coefficients and the variances are also computed using its formulation.

Conclusions

Except for Hong Kong, the three latent construct of the ARM account for approximately 20 % of the variance of student's civic knowledge. The students' perceptions of government and law and students' perceptions of public service in Asia have positive prediction effect for student's civic knowledge. While the interesting thing is that the second latent variable—students' perceptions of identity, citizenship, and culture in Asia has negative effect for student's civic knowledge.

No. PROMS2012HK004

Implementing the DIF-Free-Then-DIF Strategy on DIF Assessment

Presenter **August 7th (13:00–13:25)**

Kun Xu*, Wen-Chung Wang and Magdalena Mo Ching Mok
The Hong Kong Institute of Education
e-mail*: xukunjacob@gmail.com

Background

A prerequisite of DIF assessment is a pure scale on which the performances of reference and focal groups can be compared. In practice, all items in a test have to be assessed for DIF, such that a pure scale is not feasible before DIF assessment. To tackle this question, the DIF-free-then-DIF strategy was proposed, where a set of DIF-free items are selected first and then all other items are assessed for DIF using the DIF-free items as anchored.

Aims and Keywords

This study aims to implement the strategy on three popular DIF detection methods: the Mantel-Haenszel method, the logistic regression method, and Raju's area method, and compared their performances with traditional approaches.

Keywords Differential item functioning • Rasch measurement • Anchor item • Mantel-Haenszel method • Logistic regression method • Area method

Methods

Monte Carlo simulations were conducted. The five independent variables were: (a) method (Mantel-Haenszel, logistic regression, and area methods); (b) scale purification procedures: yes and no; (c) difference in mean ability between groups: 0 and 1; (d) percentage of DIF items in the test: 10, 20, 30 and 40 %; and (e) number of anchored items: 4 and 8. A total of 100 replications were made in each condition. The data were generated from the Rasch model and the DIF was uniform. The Type I error rate and power were computed.

Results

The traditional methods without scale purification procedures worked fairly well only when the percentage of DIF items was as small as 10 %. When scale purification procedures were implemented, the Type I error rates were well-controlled when the percentage was no greater than 30 %. The selection of a set of DIF-free items was very accurate, but not perfect. The DIF-free-then-DIF strategy was effective only when the percentage of DIF items was as high as 30 or 40 %.

Conclusion

Scale purification procedures should be implemented. The new strategy is helpful when tests have a high percentage of DIF items.

No. PROMS2012HK005

Differences in the Perception of Feedback Between Hong Kong Primary School Students and Their Teachers

Presenter **August 6th (13:30–13:55)**

Michael Ying Wah Wong[1], and Magdalena Mo Ching Mok[1,2]
[1]Assessment Research Centre, The Hong Kong Institute of Education
e-mail: mywwong@ied.edu.hk
[2]Psychological Studies, The Hong Kong Institute of Education
e-mail: mmcmok@ied.edu.hk

Background

Abundant research has been undertaken on the importance of teachers' feedback to the school achievement of students. However, focus has rarely been placed on the differences between the perception on feedback of students and teachers. Being a part of the study which was sponsored by a General Research Grant awarded by Research Grants Council of UGC, Hong Kong (Project Number 844011), we attempted to find out the differences between students' and teachers' perception on feedback.

Aims and Keywords

The study aimed to find out the differences between Hong Kong primary school students and teachers in their views to feedback.

Keywords Rasch model • Effect size • Feedback • Primary student • Primary school teacher

Sample

Four thousand five hundred and seven students between primary 3 and primary 5, 132 teachers from 26 Hong Kong primary schools were sampled in the study.

Method

A 42-item instrument with 4-Likert point response scale was constructed by the research team to collect the responses from Hong Kong students and teachers of their perception on 6 dimensions in feedback. Multidimensional Rasch analysis was conducted on the collected data with CONQUEST 2.0.

Results

The results showed that: (1) item-fit was between 0.5 and 1.5 for all items except one; (2) Rasch person reliabilities in the 6 dimensions ranged from 0.72 to 0.86 and item reliabilities was 1.00; (3) the effect size of the respondents' status (teacher or student) in the 6 dimensions ranged from 0.32 to 0.64.

Conclusions

In this study, there are moderate differences between the students and teachers in the perception on the source of feedback, the effectiveness of feedback and the function of feedback. On the other hand, there are little differences among the two types of respondents in the perception on the goodness of feedback, the focus/target of feedback and the expectation of feedback.

No. PROMS2012HK006

Facets Modeling for Rater Sensitivity

Presenter **August 6th (13:30–13:55)**
Wang Wen-Chung, Jin Kuan-Yu
The Hong Kong Institute of Education
Hong Kong

Background
Sensitivity and threshold are two important elements in making judgment (Jackson 1972). When raters are recruited to mark responses to constructed-response items, the effect induced by raters should be fully considered. However, the standard facets model (Linacre 1989) or generalized facets model (Wang and Liu 2007) accounts for the threshold element (rater severity) and does not consider the sensitivity element.

Aims and Keywords
Rater sensitivity can be defined as the effectiveness of a rater in differentiating ratees with varies degree of proficiency. By treating the combination of an item and a rater as a pseudo-item, we intend to decompose the attached slope (discrimination) parameter on the pseudo-item into two parts: item discrimination and rater sensitivity.

Sample
A dataset gathered by Congdon and McQueen (1997) was analyzed, in which each of the 8,296 students' writing scripts was graded by two raters randomly chosen from a set of 16 raters on two criteria (items) against a six-point scale.

Methods
Different formulations of facets models were fitted by using the freeware WinBUGS.

Results
The results supported our expectation: the two rating criteria had degrees of discrimination power; the 16 raters had different sensitivities; and the slope of the pseudo-items was approximately the product of item discrimination and rater sensitivity.

Conclusions
This resulting generalized facets model can account for item characteristics (difficulty and discrimination) and rater properties (severity and sensitivity) simultaneously.

No. PROMS2012HK007

Construction of the Structural Q-Matrix of the Tactics Used in Badminton Singles Games

Presenter **August 8th (11:50–12:15)**
Henry Hoi-Wai Wong
The Hong Kong Institute of Education

Background
Mr Wong Hoi Wai is interested in the researches of cognitive measurement in physical education, by implementing various cognitive diagnosis model such as DINA model, Fusion model and etc. Mr Wong is now a PE teacher in Hong Kong C.C.C. Kei Wai Primary School and concurrently a doctoral student in the Hong Kong Institute of Education.

Aims and Keywords
The main objective of this research is to explore the underlying attributes of the tactics that are frequently used by the world-class badminton singles players. This study is going to construct a set of structured attribute patterns (Q-matrix) associated with different components from a cognitive model of task performance (Linear Logistic Latent Trait Model).

Sample
This research is based on the videotaped badminton singles games held in BWF World Championship 2011. There are in total 16 games, including Quarter-Final, Semi-Final and Final Men's and Women's Single games.

Methods
Though the notational analysis on all badminton singles games and the extensive literature reviews on badminton tactics, all possible tactics will be tagged with specific attributes. For each test item, all the prerequisite information, including the positions of the opponent and the player, the velocity of the shuttlecock and the flight height of the shuttlecock will be noted. The skills that the player used will be recorded with labeling "1" for presence or "0" for absence. It is hypothesized that a particular response pattern is correspond to a particular underlying tactics. The structural matrix will be entered for analysis by the software R with extended Rasch modeling (eRm). Refining the structural matrix is needed until the goodness-of-fit is obtained.

Results
Analyses are underway. The initial result will be obtained by the end of April. The final result will be presented at PROMS conference. The initial result will be sent to PROMS once obtained.

Conclusions

Physical Education has become emphasized the tactics teaching in lessons. However, there are no substantial assessment methods for tactic-based teaching methodology. Once the ideal response patterns of tactics used by the badminton singles players are constructed, PE teachers and Sport coaches can make the diagnosis on the used tactics by using the DINA model to compare the ideal response patterns and the observed response patterns. Coaches can then monitor the competition process of player and provide the specific feedbacks to players with full substantial reasons. Players can learn from this to self-regulate their skill uses during the badminton singles games.

PROMS2012HK008

The Facets Graded Response Model

Presenter **August 8th (11:00–11:25)**
Joseph Chow; Kuan-Yu, JIN; Wen-Chung, Wang
The Hong Kong Institute of Education,
HKSAR, China
e-mail: chowkf@ied.edu.hk

Background

Raters are often recruited to mark constructed-response items. For example, student essays are graded by teachers; or teachers' performances are judged by students. Linacre (1989) has developed the facets model to account for the joint effects of item, person, and rater on item responses using the conditional-probability formulation. In the statistical literature, the latent-response formulation is used more often. The latent-response formulation is in line with the cumulative-logit models, which incorporate the graded response model (Samejima 1969) and the 2-parameter logistic model (Birnbaum 1968) as special cases.

Aims and Keywords

In this study, we aim to extend the two-facet graded response model to more than two facets so that rater effects can be evaluated. Simulations were conducted to evaluate its parameter recovery using the MPlus program.

Keywords Graded response model • Facets model • Cumulative-logit models • Parameter recovery • Mplus

Methods

The item responses were generated under the three-facet graded response model, where 5 raters judged 200 examinees on 5 tasks on a five-point rating scale. A total of 100 replications were made. The true model was used to analyze the data using the Mplus. The bias and root mean square error (RMSE) were computed to evaluate parameter recovery.

Results

The bias values were ($-0.154 \sim 0.186$), and the RMSE values were ($0.043 \sim 0.215$). Apparently, the parameter recovery of the three-facet graded response model was satisfactory.

Conclusion

The two-facet graded response model is extended to facets graded response model to examine rater effects on open-ended items. The parameters can be recovered fairly well using the Mplus. Applications of the new model are welcome.

PROMS2012HK009

Using Combining Software to Investigate the Relationship Between Rated and Counted Features of Spoken and Written Performance in English Language

Presenter **August 8th (11:25–11:50)**
Dr Cheung Kwai Mun Amy
Manager of Hong Kong Examinations and Assessment Authority (HKEAA)

Ms LEUNG Fung Yin Flora
Senior Curriculum Development Officer of Education Bureau(EDB), Hong Kong

Background
This study attempted to analyse students' performance in speaking and writing at Secondary 3 (Grade 9) in Hong Kong.

Aims
This study aimed at using combining software to investigate the relationship between rated and counted features of spoken and written performance in English Language.

Sample
A stratified sample of 180 students.

Methods

- students' performances rated by teacher raters and also subjected to verifiable quantitative measures (VQM) (Cheung 2010)
- obtained 'fair average scores from rater' ratings using *FACETS* (Linacre 1991–2008)
- calculated the VQM using *vocd* (Malvern et al. 2004) to measure 'lexical diversity (D)', and *RANGE* (Heatley et al. 2002) to measure 'types', 'families', and 'tokens'
- compared ratings against each other and against VQM to ascertain the reliability and validity of ratings
- key questions: (i) To what extent do the sub-constructs within a skill and across two skills correlate? (ii) Do students with strong written vocabulary exhibit strong spoken vocabulary, and vice versa?

Results
The main findings showed 'high' correlation levels in 'vocabulary and language patterns' across spoken and written skills.

Conclusion

Transference from writing to speaking seems more likely given the prioritization of the written mode in the Hong Kong school system and the fact that Hong Kong students usually do not have much chance to practise oral English outside school. The next phase of the study would be to develop corpus-based human assisted error identification and classification system to address learning needs in productive skills.

PROMS2012HK010

Effect of Academic Aims, Goal Setting and Planning on Academic Achievement of Secondary Students in Hong Kong

Presenter **August 8th (9:25–9:50)**

Mr. Jinxin Zhu
Assessment Research Centre, Hong Kong Institute of Education
Hong Kong
e-mail: jxzhu@s.ied.edu.hk

Prof. Magdalena Mo Ching MOK
Chair Professor, Psychological Studies Department
Hong Kong Institute of Education
Hong Kong
e-mail: mmcmok@ied.edu.hk

Background

This study is part of a larger study entitled "Secondary Students' Independent Learning" funded by a Competitive Earmarked Research Grant (Number HKIED 8005/03H) of the Research Grants Committee, Hong Kong SAR Government. The original project aimed to establish and validate with empirical data, a conceptual framework of self-directed learning, the characteristics and processes involved and their relations with academic achievement of secondary students.

Aims and Keywords

This study aimed to explore the effect of academic aims, goal setting and planning on academic achievement of secondary students.

Keywords Rasch measurement • Academic aims • Academic goal setting • Academic planning • Academic achievement • Secondary students

Sample

The sample comprised 14,846 students currently enrolled at Secondary 1 to Secondary 6 (except Secondary 5) from 23 secondary schools in Hong Kong.

Methods

Students completed a self-administered questionnaire comprising 4-point Likert items (1: Strongly Disagree, 2: Partly Disagree; 3: Partly Agree, 4: Strongly Agree) on academic aims, goal setting, planning, and other aspects of self-directed learning not included in this study. Each construct was measured by five items. Measurement quality of the scales according to the framework was established using the

Winsteps software (version 3.72.3). Based on Rasch scores of the latent variables and standardized scores of Chinese, English and mathematics, structural equation modeling was undertaken using Mplus (version 6).

Results

Analyses indicated that academic aims and academic planning both directly predicted academic achievement, with educational aims being the stronger predictor. Academic goal setting indirectly predicted academic achievement via academic planning. Consistency was found both in the structure and in the effects of the predictor across gender and year levels.

Conclusion

The study shows that self-directed learning has positive effect on academic achievement of secondary students of both gender and all year levels.

PROMS2012HK011

Rasch Validation of a Combined Measure of Basic and Extended Daily Life Functioning After Stroke

Presenter **August 6th (15:30–15:55)**

Hui-Fang Chen
Post-doctoral Fellow, Assessment Research Centre
Hong Kong Institute of Education, Hong Kong
e-mail: hfchen@ied.edu.hk

Ching-yi WU
Professor, Chang Gung University, Taoyuan, Taiwan
e-mail: cywu@mail.cgu.edu.tw

Keh-chung LIN
Professor, National Taiwan University, Taipei, Taiwan
e-mail: Kehchunglin@ntu.edu.tw

Background

Tools used to measure poststroke functional status must include basic and instrumental activities of daily living (ADL) and reflect both the patient's and the clinician's perspectives.

Aims and Keywords

This study combined the Functional Independence Measure (FIM) and the Nottingham Extended Activities of Daily Living (NEADL) to create a scale providing objective and subjective evaluations of ADL function after stroke.

Keywords Activities of daily living • Psychometrics • Rasch analysis

Samples

A total of 188 participants, who completed the NEADL and the FIM, was included in the present study. Those patients met the inclusion criteria: (1) first-ever stroke, (2) Brunnstrom stage II or above for the proximal and distal upper extremity, (3) no severe physical conditions and medical problems, (4) no cognitive impairment, and (5) no excessive spasticity at any joint of the arm.

Methods

Rasch analysis was conducted to investigate the psychometric properties of the new scale.

Results

A 3-point and a dichotomous scale were suggested for use in the FIM and the NEADL, respectively. The combined 40 items worked consistently to reflect a single construct, and "bladder management" and "bowel management" were highly related. After "bowel management" was removed, all but 3 items fit the model's expectations and showed reasonable item difficulty hierarchy with high reliability. However, the 3 misfit items were removed, and no differences were found between the 36-item and 39-item scales.

Conclusions

The combined measure provides a comprehensive picture of ADL from patients' and clinicians' perspectives. It extends the utility of the FIM and the NEADL and is recommended for use to measure independence of stroke patients.

PROMS2012HK012

Developing the Content of Certification Programme for Infection Control Nurses (ICNs) in Hong Kong

Presenter **August 6th (16:20–16:45)**

Chan WF
Advanced Practice Nurse, Infection Control Team
Tung Wah Eastern Hospital, Hong Kong SAR
e-mail: chanwf3@ha.org.hk

Adamson B
Head, Department of International Education and Lifelong Learning,
The Hong Kong Institute of Education, Hong Kong SAR
Bond T, Adjunct Professor, School of Education, James Cook University, Australia
CHUNG JWY, Peter TC Lee Chair Professor of Health Studies,
Department of Health and Physical Education, The Hong Kong
Institute of Education, Hong Kong SAR
CHOW MCM, Assistant Professor, School of Nursing,
The Hong Kong Polytechnic University, Hong Kong SAR

Background

Nursing specialization is the world trend in healthcare development. Certification is one of the means to recognise the nurse specialists. There is no certification system in nursing field in Hong Kong. Participating in the certification examination in the overseas does not address the problem because the practices of nurse specialists are local-context specific.

Aims and Keywords

The aim of this research is to develop the content of certification programme for ICNs in Hong Kong.

Keywords Infection control nurse • Certification • Hong Kong

Sample

Expert panels, 18 ex-ICNs and 117 ICNs in Hong Kong

Methods

This is a 3-phase research. Phase One proposed the core competency of ICNs by literature review and Delphi survey. Validity and reliability were established employing content validity survey by experts and repeated surveys on ex-ICNs. Phase Two confirmed the core competency of ICNs through an opinion survey on practicing ICNs. Data were analysed using Winsteps programme. A qualitative questionnaire survey on experts was conducted to define the critical competency of

ICNs in Phase Three. With Rasch-based concept, safety margin was added as resulted in true critical competency. The items were rescaled and importance levels (in logit) were transferred to content weight to develop the content blueprint of certification programme for ICNs of Hong Kong.

Results

The core competency of ICNs identified in Phase Two consisted of 76 items with importance levels (in logit) attached in individual items. The competency scale was unidimensional with very good reliability estimates. The critical competency was the most important portion of the core competency. The true critical competency comprised 35 items, including 25-item expert-defined critical competency and 10-item safety margin.

Conclusions

The Rasch analysis concluded the core competency of ICNs in Hong Kong by identifying the items objectively and deciding the respective importance levels. Based on the expert-defined critical competency, Rasch-based concept was employed to ensure the inclusion of essential items in the critical competency that serving as the content blueprint of certification programme for ICNs. This is the first reported work on developing the content of certification programme for nurse specialists in Hong Kong. This study can be replicated in other nursing specialties for developing the content of respective certification programmes. It also served as the ground work for nursing professional development in future.

PROMS2012HK014

Applying the Rasch Model to Diagnostic English Language Testing

Presenter **August 7th (16:25–16:50)**
Jerry Gray[1], Gwendoline Guan[2], (email: ygguan@cityu.edu.hk)
Michelle Raquel[3], Winnie Shum[1], Carrie Tsang[3], Alan Urmston[3], Roxanne Wong[2]
[1]Lingnan University
[2]City University of Hong Kong
[3]The Hong Kong Polytechnic University

Background

In 2009, three Hong Kong institutions shared their expertise and resources for the purpose of developing and establishing a diagnostic assessment to enhance students' English language learning. They aimed to develop an instrument of the highest quality in terms of reliability, validity and usefulness. Based on Rasch measurement principles, they developed the Diagnostic English Language Tracking Assessment (DELTA). Through the DELTA, students can diagnose their strengths and weaknesses in reading, listening, grammar and vocabulary; track their English language gains; and plan their English language learning.

Aims and Keywords

This research aims to investigate the application of Rasch measurement to the diagnosis of English language proficiency.

Sample

Two thousand and five hundred Year one undergraduates, the DELTA.

Methods

Rasch measurement, Winsteps software package.

Results

- Calibrated person measures to track growth over time;
- DIF analysis to demonstrate students' strengths and weaknesses on reading, listening, vocabulary and grammar;
- An adapted KIDMAP to show individual performance on a set of items on each component of the test

Conclusions

The DELTA test is a valid and reliable instrument that measures students' English language proficiency, and the Rasch Model can be applied to diagnostic language testing.

PROMS2012HK015

Constructing Rasch-Based Measures for Assessing Academic Attainment for Students with Intellectual Disabilities

Presenter **August 7th (13:25–13:50)**

Vicky TSang; Trevor G. Bond; Joseph Chow
The Hong Kong Institute of Education, HKSAR, China
James Cook University, Australia
e-mail: chowkf@ied.edu.hk

Keywords Rasch measurement • Scale construction • Students with intellectual disabilities • Curriculum-based assessment • Evaluation tool

Background

Since 2005, professionals in nine special schools collectively had developed an assessment tool called SCALES for all students regardless they are in special or mainstream schools in Hong Kong. Based on the concept of "one common curriculum for all", SCALES is classified into the four Key Learning Area (KLA) strands of learning as in the mainstream curriculum framework, namely, Chinese, Science, Personal, Social and Humanities Education (PSHE) and Mathematics. This means that for the first time all of the students are being measured by exactly the same criteria irrespective of the ability or disability.

Aims

In this study, we aim to apply Rasch measurement approach to analyzing a this innovative evaluation tool for children with special needs due to intellectual disabilities. Since Rasch measurement can convert ordinal measures into interval measures, the latter of which is fundamental for objective measurement in human sciences (Bond and Fox 2007).

Methods

By using Winsteps, statistics such as infit and outfit mean square, person separation index, reliability, and dimensionality of the measurement scales are supported empirically.

Results

The uni-dimensionality of the four major subject areas is supported with good psychometric properties such as high reliability and person separations that help differentiation of subjects. Selective applications of some items of these scales could be used by practitioners for quick screening and better intervention for care of children with special leaning needs.

Conclusion

The scales are supported with both subject experts' observation and quantitative evidence from sophisticated psychometric tests. Successful establishment of the instrument by subject experts is evident and quality-assured administration of the Rasch-based items in the future is recommended as the SCALES has proved to be a valid and reliable assessment measure for diverse learners.

PROMS2012HK016

Identification of the Patterns of Chinese Character Recognition in Students with Learning Disabilities Requiring Tier-2 Support: A Rasch Analysis

Presenter **August 7th (15:35–16:00)**
Fuk-chuen HO and Zi, Yan
Hong Kong Institute of Education

Background
Dr Fuk-chuen Ho is Assistant Professor in the Department of Special Education and Counselling and the associate director in the Centre for Special Needs and Studies in Inclusive Education at the Hong Kong Institute of Education. He was formerly an inspector in the Special Education Inspectorate of the Hong Kong Education Bureau. He is now the project leader of three external funded projects in the areas of dyslexia, Theory of Mind and collaborative mode of professional development for teachers in special schools respectively. (fcho@ied.edu.hk)

Dr Zi Yan is Assistant Professor in the Department of Curriculum and Instruction at the Hong Kong Institute of Education. He has a research interest in educational measurement and inclusive education. He is now participating in a commissioned project on the evaluation of the inclusive education in Macau. (zyan@ied.edu.hk)

Aims and Keywords
This study investigates the Chinese reading patterns of students with learning disabilities (LD). The performances of students with LD in reading the three categories of Chinese characters were particularly analyzed: regular, irregular, and pseudo-characters.

Keywords Learning disabilities • Chinese character recognition • Regularity

Sample
Fifty-three students with LD in reading and 44 average students at age 9 were selected from five Hong Kong primary schools.

Method
Their abilities for reading Chinese characters were measured using Rasch analysis.

Results
Both types of students found regular characters as the easiest to read. Average students showed better performance in reading irregular characters than pseudo-

characters, whereas students with LD exhibited no significant performance difference in reading these two categories.

Conclusions
Students with learning disabilities and students without disabilities have different preferences for the phonological and orthographic routes to read Chinese characters.

PROMS2012HK017

Applying the Many-Facet Rasch Model to an Exit English Speaking Proficiency Assessment

Presenter **August 6th (13:55–14:20)**
Felicia Fang, Alan Urmston
The Hong Kong Polytechnic University
e-mail: Alan.Urmston@inet.polyu.edu.hk

Aims and Keywords
This research aims to investigate the application of Many-facet Rasch measurement in the reporting of test scores to students and the university, the analysis of the performance of the tasks and criteria and the performance of raters in a university-level exit English assessment of speaking in Hong Kong.

Sample
Approx. 3,000 students who sat the assessment in academic year 2011–2012, 30+ raters, 5 tasks.

Methods
Many-facet Rasch analysis was conducted using the FACETS Programme (Linacre and Wright 1993).

Results
The major findings of the study:

- FACETS proved to be a reliable instrument in providing a clear picture of how students performed on this test, how well the tasks performed, and how reliable the raters were when rating the performances
- The all facet vertical rulers
- Candidate measurement report: candidates at different levels were clearly distinguished
- Rater measurement report: misfitting (inconsistent, and/or overly harsh/lenient) raters were identified
- Task measurement report: how well the tasks performed
- Item measurement report: how well each criteria on each task performed
- The separate rater report: provided information on rating consistency, the rater's individual use of the levels (e.g. whether clear discrimination between levels was achieved, whether any particular level was under representation, etc.)

Conclusions
The results informed future rater accreditation and training, and the writing of new test tasks.

PROMS2012IR001

Implication of Differential Item Functioning in Investigating the Possible Effect of learners' Identity on the Underlying Structure of Listening Comprehension Ability

Presenter
Parisa Daftarifard
A PhD candidate at Azad University (Science and Research Branch)
Tehran, Iran
e-mail: pdaftaryfard@azad.ac.ir

Minoo Alemi
A PhD candidate of Applied Linguistics at Allameh Tabataba'i University and a faculty member of Languages and Linguistics Department at Sharif University of Technology, Iran
e-mail: alemi@sharif.ir

Background

Second Language (L2) Self identity gains a focal attention in research enquiries nowadays (for example, LoCastro 2001) though filtering hypothesis (Krashen 1981) and research on good language learner (GLL) specifically when culture is the main focus (Finkbeiner 2008) and when belief is attributed to GLL (White C 2008). We hypothesized that items of listening comprhension might function differently in terms of learners' level of identity. To test this, one way is to use the method of differential item functioning (DIF). DIF studies mostly focus on age, gender (Maller 2001) different background knowledge (Pae 2004), different first languages (Ryan and Bachman 1922) but to our knowledge, there is not much research done on investigating the effect of different identity type on test performance. To this end, the following questions were posited.

- What is the relationship between L2 self identity and different levels of listening ability?
- Does L2 self identity can predict learners' score on different listening sub-skills?
- Are there some different target populations in terms of identity in the focal group?

Sample

About 60 Iranian students of Engineering at Sharif University participated in this study.

Method

The method of this research is ex post facto because no treatment is provided and no manipulation is done. The data is gathered through two instruments: the listening

section of TOEFL and a questionnaire of self identity adapted from LoCasro (2001). Data is analyzed through both Classical True score theory and Rasch model using Winstep.

Result

The result revealed that not all sub skills of listening comprehension are affected by Iranian learners' identity.

No. PROMS2012JP002

Learning by Assessing in Second Language Writing

Presenter **August 7th (15:35–16:00)**
Trevor A. Holster
Fukuoka Women's University
e-mail: trevholster@gmail.com

Bill Pellowe
Kinki University
e-mail: pellowe@fuk.kindai.ac.jp

J. Lake
Fukuoka Women's University
e-mail: jlake@temple.edu

Aaron Hahn
Fukuoka High-school
e-mail: qwyrxian@hotmail.com

Background
Recent studies have used many-faceted-Rasch-measurement (MFRM) to investigate the potential of peer assessment (PA) in second language classrooms. However, "learning by assessing" (LBA), where the assessment process results in learning by raters, has not been widely investigated.

Aims
This study aimed to demonstrate LBA in second language writing by testing the research hypothesis that gains measured on a PA rubric would exceed gains on a secondary rubric measuring general academic writing.

Keywords Peer assessment • Learning by assessing • Writing proficiency • Second language assessment

Sample
A convenience sampling of 26 first year university students in an academic English program participated as part of a compulsory academic writing course.

Method
Pre and post treatment writing was rated by teachers on the rubric used for PA and a general academic writing rubric, using the *Facets* software package. Students were exposed to the PA rubric, but not to the secondary rubric. The *MOARS* audience response system allowed students to rate each others' work and to access their own ratings instantaneously using internet connected mobile devices.

Results

Substantive gains were observed on the PA rubric but not on the secondary rubric, providing evidence of LBA. Response patterns suggested holistic rating by peer raters, casting doubt on the potential of PA to provide formative feedback and supporting the view that the gains observed arose from LBA.

Conclusions

The holistic response patterns displayed by peer raters discounted PA as a source of diagnostic feedback. Instead, the results suggest that the major benefit of PA lies in improving peer raters' knowledge of the target construct through repeated engagement with the rubric during the rating process.

No. PROMS2012JP003

Metalinguistic Knowledge of Low-Proficiency EFL Learners at a Japanese University

Presenter **August 7th (13:50–15:15)**
Miki Tokunaga
Fukuoka University
e-mail: tokunagamiki@gmail.com

Background

Recent studies have shown correlations between learners' L2 proficiency and their L2 metalinguistic knowledge. Although many of the studies were conducted on learners with intermediate or higher levels of proficiency, a much simpler test with only high frequency words is necessary to measure metalinguistic knowledge of low proficiency EFL learners.

Aims and Keywords

This research aimed to design a simple metalinguistic test to find out what metalinguistic features are recognized by low proficiency EFL students and whether there is a correlation between their English proficiency and metalinguistic knowledge.

Keywords Metalinguistic knowledge • Explicit knowledge • Remedial education

Sample

The participants were 708 mixed major students taking required English classes at a private university in Japan.

Methods

A metalinguistic test was designed using simple L2 vocabulary to minimize the effect of learners' vocabulary knowledge. The items aimed to test basic metalinguistic knowledge by identifying parts of speech, parts of sentence patterns, tenses and word structures. Winsteps was used to analyze the learners and items and results were compared with proficiency test (TOEIC Bridge) results.

Results

Moderate correlations were found between the Rasch measures of the metalinguistic knowledge and all sections of the TOEIC Bridge scores. The strongest correlations was with the Reading section ($r = .66$). Many of the participants had difficulty identifying simple metalinguistic features.

Conclusions

The results suggest that many of the subjects have insufficient metalinguistic knowledge to understand explanations by teachers and textbooks. Many universities in Japan now offer remedial level English classes teaching basic grammar and vocabulary. However, simplified content alone does not solve the problem when learners cannot understand instructions. Understanding students' metalinguistic knowledge and providing targeted remedial instruction is thus necessary.

No. PROMS2012JP004

Mobile Audience Response System for Peer Assessment

Presenter
Bill Pellowe
Kinki University
e-mail: pellowe@fuk.kindai.ac.jp

Trevor A. Holster
Fukuoka Women's University
e-mail: trevholster@gmail.com

August 7th (14:15–14:40)

Background
Previous studies have investigated peer assessment (PA) in second language classrooms, but the very large datasets generated by PA limit its practicality for regular classroom use. However, internet-connected mobile devices allow students to enter PA ratings directly into an online database, allowing for immediate analysis and feedback in the classroom.

Aims
This project aimed to develop a peer assessment module for the open source *Mobile Audience Response System* (*MOARS*), allowing users to enter ratings using any mobile device with a wireless internet connection. The *MOARS* PA module provides output in the form of a specification file and data file formatted for immediate analysis using the *Facets* or *Minifacs* software packages, allowing users with minimal technical skills to conduct MFRM analysis.

Sample
This pilot study comprised approximately 170 first year students enrolled in presentation skills classes in an academic English program at a Japanese university.

Method
The *MOARS* PA module was tested operationally. Students and teachers rated training videos, providing linking data to equate disjoint class groups. Students then rated each others' live presentations in class. Data was collected using *MOARS* and analyzed using *Facets*, providing equated ability measures for all students in the 12 class groups and diagnostic analysis of students and teacher raters.

Results
No major problems were encountered with the *MOARS* system. Communicating diagnostic results to teachers and students was the major difficulty, requiring development of graphic outputs to supplement those available from *Facets*.

Conclusions

The combination of *MOARS* and *Facets* resolved problems of the practicality of collection and analysis of PA data in the classroom. The next stage of development will focus on simple graphical outputs to simplify diagnostic analysis for classroom teachers.

No. PROMS2012JP005

Can Difficulty of Items Be Guessed Intelligently Without Degrading CAT Results?

Presenter
Tetsuo Kimura
Niigata Seiryo University
e-mail: tetsuo.kmr@gmail.com

Keizo Nagaoka
Waseda University
e-mail; k.nagaoka@waseda.jp

August 8th (9:00–9:25)

Background

The authors constructed item banks for in-house computer adaptive tests (CATs) for EFL learners (Kimura 2009; Kimura and Nagaoka 2010). After several pretesting and equating by common item method, the number of item reached 258 for vocabulary and grammar and 307 for listening comprehension. The Rasch-based computer adaptive test program named UCAT written in BASIC (Linacre 1987) was converted into PHP so that CATs can be administered on a major open source learning management system (Kimura, Ohnishi and Nagaoka 2012).

Aims, Sample & Method

Linacre (2000) argued that the difficulty level of the new items can be guessed intelligently without degrading the resulting ability estimates in CAT. The aim of this study is to confirm this argument by comparing the results of two CATs administered to the same test takers (59 Japanese freshmen of engineering department). Both CATs used the same 258 items for vocabulary and grammar. The difficulty level of items of the first CAT was determined by pretesting ($M = 0.19$, $SD = 1.37$). That of the second CAT was guessed intelligently either one of four levels (-1.5, -0.5, 0.5, or 1.5).

In either test, the initial estimate ability was set as 0.0 and 16 items were selected adaptively. The first item was selected randomly between -0.5 and 0.5. The next item was selected randomly between lower limit (LL: ability estimate when the answer would be wrong) and upper limit (UL: ability estimate when the answer would be correct). If no item was found between LL & UL, the closest one was selected.

Keywords Computer adaptive test • Item difficulty • Rasch

Results & Conclusions

The ability estimate averages of the two CATs were not significantly different (0.81 and 0.66). The standard errors were almost identical (0.53 and 0.54). The correlation coefficient between the two measurements was high (0.83). Consequently the current study suggests that the difficulty of items can be guessed intelligently without degrading the CAT results.

No. PROMS2012JP006

Abstract Proposal for PROMS 2012

Presenter
Aaron Olaf Batty
A visiting assistant professor at Keio University's Shonan Fujisawa Campus, in the Department of Environment and Information Studies.

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 and Keywords
The present research investigates the interactions between level, accuracy, and confidence with regards to special uses of the utterance verbs "speak," "talk," "say," and "tell."

Keywords MFRM • English linguistics • Accuracy vs. 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 classtime 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. These data were scaled in Facets (Linacre, *Facets*. Beaverton: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm, 2011) and compared via the methods developed by Paek et al. (*A study of confidence and accuracy using the Rasch modeling procedures* (*Research Report No. RR-08-42*). Princeton: Educational Testing Service, 2008).

Results
Overall, even high-level learners of English demonstrated ignorance of many of the common constructions tested. The verb "talk" proved especially difficult, even for non-native speakers who had spent years living in Anglophone countries. Generally, confidence tracked accuracy weakly.

Conclusions

Some core meanings of verbs in English, and the "exceptional" uses they cause, may need to be explicitly taught, as it seems that learners do not naturally acquire these, even with a high degree of exposure to native speakers.

References

Linacre, J. M. (2011). *Facets*. Beaverton: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm.

Paek, I., Lee, J., Stankov, L., & Wilson, M. (2008). *A study of confidence and accuracy using the Rasch modeling procedures* (*Research Report No. RR-08-42*). Princeton: Educational Testing Service.

PROMS2012JP007

Using Student Experience for Class Composition

Presenter **August 7th (16:00–16:25)**
Jeffrey Durand and Forrest Nelson
Associate Professors at Tokai University
Japan
e-mail: kandajeffd@gmail.com

At this large Japanese university, third-year students were streamed into an English class according to weighted test scores, with some difficulty from missing scores. For the most part, this method produced reasonably homogeneous classrooms according to ability, but two to five students sometimes appeared to belong in a different level. The purpose of this research is to develop a method of better separating students for purposes of class placement. Accordingly, information on students' study experiences prior to university was used. The goal is to quickly obtain relevant, useful information to better construct classes. All students from the third-year course were used. A small number of questionnaires were not returned, mainly from students who never attended the class. Students were placed into classes according to test scores. On the first day of class, they completed a questionnaire on their experiences studying English. Winsteps was used to fine-tune the questionnaire and obtain experience scores for each student. Winsteps was also used to both compare and combine the experience items with the test scores. In addition, each teacher responded to a questionnaire during the first class and later in the year on whether students were appropriately placed. In fine-tuning the questionnaire, some kinds of experience "did not matter" in connection with student ability. More importantly, the experience scores improved student separation for streaming purposes. However, questions remain about the unidimensionality of this 'ability to succeed' construct. Using information on student experience in a Rasch analysis can greatly improve student placement. The unidimensionality issue has brought up a broader issue, however, of what the goals of class streaming should be: what kind of homogeneity is desired.

No. PROMS2012MS001

Using Rasch Model Analysis to Investigate and Compare Teachers' Conception of Formative Assessment

Presenter **August 6th (14:45–15:10)**
Sharifah Norsana Bt Syed Abdullah
e-mail: eshalbiha@gmail.com

Mohamed Najib B Abdul Ghaffar
e-mail: p-najib@utm.my

Malaysia has been hard at work to transform its education system. In line with the Education Transformation Programme (ETP) to empower the education system, government had introduced a new assessment system which is more focus on outcome-based education and not be too examination oriented. A movement in assessment paradigms from measuring the amount of learning to enhancing learning which focused on more contextualized, communicative, performance-based as well as authentic assessment. Therefore, in this new assessment system more emphasis will give to classroom assessment rather than standardize test. Teachers must be prepared to integrated assessment with instruction in classroom. Teacher assessment plays a useful role in any assessment system because it will provide specific feedback on the progress of individual pupils to feed into teaching and learning and informs about individual strengths and weaknesses.

A descriptive study was conducted to identify the conception of formative assessment among teachers in Gombak district, Selangor. The main objectives of this study were to investigate teachers' conception of formative assessment and to compare teachers' conception in different teaching levels and subject areas. The Rasch rating scale measurement model was applied to the responses of primary, secondary and high school teachers ($N = 150$) from 6 schools in Gombak district, Selangor. Analysis of Differential Item Functioning was performed to compare the teachers' conception of formative assessment according to subject areas and teaching levels. The results show that conception of formative assessment of social science and science teachers, primary, secondary and high school teachers are different. Analysis of the data revealed that the TFAC is a valid and reliable instrument to measure teachers' formative assessment practices. The findings provided the Ministry of Education with useful information for training of school teachers in formative assessment practices in classroom.

No. PROMS2012MS002

Rasch Analysis of Goal Orientation Test in a Heterogeneous Setting

Presenter **August 8th (10:15–10:40)**

Monsurat Olusola Mosaku, and Mohd Najib Ghafar
Department of Educational Foundation
Faculty of Education
Universiti Teknologi, Malaysia
e-mail: mabaqo@ymail.com, p-najib@utm.my

Higher education providers of the knowledge based economies contend with the challenges of internalization and diversification in order to be globally relevant and competitive. Assessment in contemporary higher education is expected to measure and evaluate cognitive and non-cognitive learning outcomes as both make up the Key Performance Index used to appraise these institutions. Students consequently formulate a mechanism in order to achieve both learning outcomes. Goal orientation implying students' disposition at allocating varying amount of effort and time in order to achieve their aims is one of such mechanisms. Abundant literature asserts the implication of goal orientation in the successful learning experience and outcome of higher education. However, can it be generalized to a heterogeneous setting? This study aims to develop a valid instrument interpreting goal orientations among undergraduates in a heterogeneous setting. For this purpose, an exploratory quantitative research design using survey technique will be implemented among 254 undergraduate students in a multi stratified sampling procedure to represent the population of a university. Content, construct, predictive validities will be determined using Partial Credit Model of WINSTEP. Analysis resulted in variation based on discipline and religious beliefs while no significant effect was based on gender. This study will help to understand and appreciate the heterogeneous nature among Asian undergraduates.

Keywords Goal orientation • Higher education • Rasch analysis

No. PROMS2012MS003

Validation of Fractions Skills Test for Primary Students

Presenter
Shafiza Mohamed
Malaysian Examinations Council

Kamisah Osman
National University Of Malaysia

August 6th (15:55–16:20)

Background
Most of the students struggle to understand fractions as well as individual adults. Proficiency in fractions is crucial and useful in everyday life especially in the measurement context.

Aims and Keywords
The purposes of this study are to validate and revalidate the fractions skills test amongst the primary students.

Keywords Fractions skills • Primary students • Conceptual knowledge • Procedural knowledge • Rasch model

Sample
Total of sample was 160 students (78 boys and 82 girls) from Year Five (aged 11 years) and Year Six (aged 12 years) in Malaysia.

Methods
A fractions skills test (83 items) was administered using paper and pencil. The test comprises of conceptual knowledge construct (Part-Whole, Proportion and Number subskills) and procedural knowledge construct (Simplification, Problem Solving and Computation subskills). Rasch dichotomous model approach was applied using Winsteps (version 3.71.0.1).

Results
Result shows the high internal reliability and demonstrated a goodness-of-fit of items except for 6 items. The sample shows high internal reliability of students. Items-students mapping shows on-targeted but with a small ceiling's effect is detected.

Conclusions
This study is part of the ongoing investigation process to verify a valid test in classroom assessment for fractions' teaching and learning. Students who did not perform well on fractions skills will lead to the difficulties to learn decimals, percentages and algebra.

PROMS2012MS007

Validation of the Internet Addiction Scale in Context of Higher Education: Applying Rasch Model

Presenter **August 7th (13:25–13:50)**

A.Y.M. Atiquil Islam
Institute of Education
International Islamic University Malaysia (IIUM)
PhD Student in Education
e-mail: skyiium@yahoo.com

Muhammad Mehedi Masud
Department of Economics
University of Malaya (UM)
PhD Candidate
e-mail: mehedi_rajapur@yahoo.com

In this age of exponential knowledge growth, where internet is playing a dominant role, the authorities of Higher Education concerned have to ensure that this tool remains within the reach of the students. However, despite a decade of existence, the internet was discovered to be addicted, especially by students. In so doing, the objective of this study is to determine the extent of students' addiction in using internet and examine the validity and reliability of the Internet Addiction Scale. A total of 200 students from four faculties (Economics, Human Science, ICT and Engineering) were selected using quota sampling procedure. A questionnaire consisting of items validated from prior studies was put together and modified to suit the current study. A five-point Likert scale asking the respondents of the extent of their agreement/disagreement to the items constituting the construct in the questionnaire was used. The questionnaire's validity and reliability were established through a Rasch model using Winsteps version 3.94. The results exhibited that (i) the items reliability was found to be at 0.94 (SD = 55.7), while the persons reliability was 0.88 (SD = 12.9); (ii) the items and persons separation were 4.02 and 2.76 respectively; (iii) all the items measured in the same direction (ptmea. corr. >0.36); (iv) all items showed good item fit and constructed a continuum of increasing intensity. The findings of this study foster support for the internal consistency reliability, unidimensionality, and measurement properties of the Internet Addiction Scale which is valid.

Keywords Internet addiction • Rasch model • Higher education

PROMS2012MS008

Development and Validation of the Corporate Citizenship Scale: Applying Rasch Model

Presenter **August 7th (15:10–15:35)**

Kamala Vainy Kanapathi Pillai
Centre of Graduate Studies
Open University Malaysia (OUM)
PhD Student in Business Administration
e-mail: kamalavarny@yahoo.com

A.Y.M. Atiquil Islam
Institute of Education
International Islamic University Malaysia (IIUM)
PhD Student in Education
e-mail: skyiium@yahoo.com

Corporate citizenship practices continue to accelerate as more global social and environmental issues arise; hence leading to increased pressure on authorities, leaders and corporations in Malaysia to re-evaluate their roles and impacts on society and environment. However, critical analyses to date indicate that there has been a lack of study on corporate conduct as well as firm-stakeholder collaboration within the Malaysian context, therefore, creating a gap for investigation. In so doing, the objective of this study is to develop and validate the corporate citizen scale to measure the corporate conduct of companies in Malaysia. The data for this study was collected through an online survey administered on senior managers involved in corporate citizenship efforts. A total of 31 managers from various industries participated in this study. The instruments' reliability and validity were conducted by Rasch Model using Winsteps version 3.49. The results of Rasch Analysis demonstrate that (i) items and persons measured reliabily ($r = .79$, and $r = .96$, respectively), (ii) all valid items measured in the same direction (ptmea. corr. >0.30), (iii) most items depict good item fit and construct a continuum of increasing intensity. The findings reveal that this study's significance stands on its contribution towards developing and validating the corporate citizenship scale that could be applied by future researchers in diverse educational and industrial settings.

Keywords Rasch model • Corporate citizenship • Corporate conduct • Stakeholder coll

PROMS2012MS009

Predictors of Self-Handicapping Behaviours Among Muslim Youths

Presenter **August 8th (11:25–11:50)**
Hafsa Mzee Mwita, PhD Counseling Student
Syed Alwi Shahab, Associate Professor
Mohamad Sahari Nordin, Professor
Ainol Madziah Zubeir, Associate Professor
Mohd Burhan Ibrahim, Assistant Professor
Siti Rafiah Abd Hamid, Assistant Professor
Institute of Education, International Islamic University Malaysia, KL Malaysia

The purpose of this study is to examine the degree to which academic self-handicapping behaviour exhibited in young adults are influenced by various degrees of student engagement. One's behaviour is in accordance to ones belief about oneself (Woods 1998), thus academic self-concept is critical in the academic growth of the student because it has a direct effect on the students' performance, parents' and community expectations, student's future career as well as his/her lifestyle and successes – in this and in the after-life. Survey has been conducted to 1,032 students and the return was 832, whereby only 790 were correctly filled, thus the analysis is based on 790 respondents who are undergraduate students of International Islamic University Malaysia. Self-concept of each student was identified by utilizing self assessment tools with an aim of assessing the presence and/or the degree of students' academic self-handicapping behavior as well as their stand, commitment and engagement to the university requirements and activities. Hence, this study investigates whether self-handicapping behavior and student engagement (emotional, cognitive, behavioral and religious engagements) represents five conceptually and empirically distinct psychological constructs when studied within the same domain; the nature of relationship existing between the four inter-related constructs of student engagement; the predicting quality of the four student engagement constructs on SHB construct; the fitness of the model of predictions of self-handicapping behaviour and the moderating quality of gender and nationality on the model of predictors of self-handicapping behaviour. Results proved the four factor model of predictors of SHB as empirically fit and reliable, while the SHB tool is being studied through Rash Model so as to formulate a partial disaggregation SEM model which would consist of one latent variable and four manifest variables.

Keywords Academic self-handicapping behavior • Student engagement • Counseling psychology

PROMS2012MS010

Investigating the Possible Hierarchical Order of Reading Skills Using the Many-Facet Rasch Model

Presenter **August 6th (14:45–15:10)**
Kamal J I Badrasawi and Noor Lide Abu Kassim
e-mail: kamalbadrasawi@gmail.com

Reading ability in English as second or a foreign language is highly demanded as English has been extensively used in all fields of human knowledge. Thus, much research has been conducted to identify the nature of reading skill in L2. Two major views could be figured out: reading as a 'unitary' skill and reading as 'multi-divisible' skill. Despite this, the literature shows a lack of consensus in determining the number of the skills/sub skills that reading includes, and whether they are hierarchically ordered. Some studies have found that item difficulty is influenced by a number of factors other than person ability such as item test features i.e., the interaction between item characteristics and item difficulty. These features include, among others, text type, context type, test length, item format etc. As such, it has become problematic to ascertain the hierarchical assumption of reading skills. To resolve this, the multifaceted approach has been recommended to examine item difficulty taking into consideration factors which affect item difficulty level. This study employs the Many-faceted Rasch Model to ascertain the hierarchical assumption of reading skills for support of a developmental of reading ability. A 42-MCQ item test was administered to 944 ESL secondary students in Malaysia. The test items were identified according to the skill areas associated with them (interpreting information, making inference, understanding figurative language, drawing conclusions, scanning for details, and finding out word meanings); context type (linear and non-linear); and text type (ads, notices, a chart, a story extract, short messages, a poem, a brochure, and a formal letter). As connectivity is established for skill areas and context type only, the results for these two aspects are reported as they can be compared directly and replicated. The results show that item context types are not equally difficult; linear context types tended to be more difficult than non-linear context types. And the skill categories do not have the same difficulty level. The most difficult skill category was interpreting information and the easiest one was finding out word meanings. This supports that notion that different reading skills exert differential cognitive demands and those that require higher order thinking skills such as analysis are more difficult than those requiring lower order skills such as finding out word meaning and scanning for information. To conclude, the results of the FACETS anlyses have provided the much needed evidence that there is a strong possibility of such reading hierarchy.

Keywords Reading hierarchy • Reading skills/sub skills • Multifaceted Rasch model

PROMS2012MS011

Applying Rasch Model to Assess Interactive Whiteboard Scale

Presenter **August 7th (16:25–16:50)**

Essam Saleh Ahmed Bakadam
University Technology Malaysia (UTM)
PhD Student in Education
e-mail: essam-12@hotmail.com

A.Y.M. Atiquil Islam
Institute of Education
International Islamic University Malaysia (IIUM)
PhD Student in Education
e-mail: skyiium@yahoo.com

An interactive whiteboard (IWB) is a large, touch-sensitive board which is connected to a digital projector and a computer. The projector displays the image from the computer screen on the board. The computer can then be controlled by touching the board, either directly or with a special pen. The IWB is a useful interactive device that holds the students attention, and accommodates different learning style due to its interactive component, it also improves the way a teacher presents information, by utilizing it the Teacher will have less classroom management problems because interactive whiteboards keep students stimulated and interested. However, since its implementation at the Prince Sultan Intermediate School (PSIS) 2007, there has been no research conducted to analyze the use of the IWB by teachers, what teachers perceive to be the benefits and the problems encountered. In so doing, the aim of this study was to develop and validate the Interactive Whiteboard scale to assess the perception of the teachers on benefits in using IWB for their teaching and learning. The survey respondents consisted of 50 male teachers were selected using purposive sampling technique those who were utilizing the IWB in their instructions at the PSIS. A five-point Likert scale that indicated degrees of agreement/disagreement was used to capture the teachers' views about the benefits of the board. The questionnaire's reliability and validity were conducted through a Rasch model using Winsteps version 3.94. The results demonstrated that (i) items and persons measured reliabily ($r = .73$, and $r = .75$, respectively); (ii) all items measured in the same direction (ptmea. corr. >0.49); (iii) most items showed good item fit and construct a continuum of increasing intensity. The findings also discovered that the variance explained by the measures was 47.5 % which indicated that the items were able to endorse the PSIS teachers' benefits in using IWB.

Keywords Interactive whiteboard • Rasch model • Perceived benefits

PROMS2012MS012

Motivation and Arabic Learning Achievement: A Comparative Study Between Two Types of Gansu Islamic Schools in China

Presenter **August 7th (15:10–15:35)**

Juping Qiao
e-mail: qiaojuping@gmail.com

Prof. Noor Lide Abu Kassim
e-mail: dr.noorlide@gmail.com

Dr. Kamal Badrasawi
International Islamic University Malaysia, Malaysia
e-mail: kamalbadrasawi@gmail.com

Background
Motivation is one of the most highly studied issues within the field of L2 learning. A number of theories of motivation have been used to explain the effects of motivation on L2 learning. The Self-determination theory (SDT), a popular motivational theory developed by Ryan and Deci (1985), has been applied by many L2 researchers to explicate the relationship between motivation and L2 learning outcomes from social and psychological perspectives. In addition, SDT distinguishes intrinsic motivation from extrinsic motivation on L2 learning achievement. Bakar et al. (2010) in their study extended Ryan and Deci's self-determination theory to investigate the role of religious motivation in Arabic (L2) language learning achievement. This study sought to extend previous findings by examining the relationship between motivation aspects: Religious Motivation (RM), Internal Motivation (IM), External Motivation (EM), and Amotivation (AM) and Arabic language learning achievement.

Aims and Keywords
Specifically, it aims to (a) explore factors that influence Arabic learning achievement and examine the relative contribution of the different aspects of motivation (RM, IM, EM and AM) on Arabic language learning achievement; (b) investigate the relationship between those aspects; and (c) compare the level of motivation across two subgroups of Arabic language learners.

Keywords Self-determination theory • Arabic learning achievement • Religious motivation • Rasch analysis

Methods
To achieve these purposes, 348 students were randomly selected from two types of Gansu Islamic schools in China to complete a 36-item questionnaire.

Interviews were also conducted with four respondents to get more in-depth information as regards Arabic language learning. Qualitative analysis was conducted to answer the first research question, while the Rasch Measurement Model, independent sample t-test, Person correlations, and multiple regression analysis were utilized to answer the other research questions.

Results

The qualitative analyses indicated that there are some positive and negative factors (internal and external) that affect student learning of Arabic language either positively or negatively. Among the positive factors are religion, positive attitudes towards learning Arabic, and finding a more prestigious job. On the other hand, the negative factors include, lack of motivation to learn Arabic and the teaching methodology. The quantitative data showed significant correlations among all the motivation aspects except Amotivation. The multiple regression analyses indicated that AM and RM were significant predictors of Arabic language learning.

Conclusions

All the motivation aspects influence students' Arabic learning achievement either positively or negatively. Religion and a positive attitude towards Arabic language learning motivate students to do better in learning the language, while Amotivation and inappropriate teaching methodology deter students from learning Arabic effectively. Direct intervention and new learning and teaching strategies should be formulated to promote effective learning of Arabic language.

PROMS2012NL001

Can Mindfulness Be Measured Using Self-Report Questionnaires? A Critical Examination of the Five Factor Mindfulness Questionnaire by Rasch Model

Presenter **August 6th (14:20–14:45)**
Ying Zhang
e-mail: Ying.Zhang@umcg.nl

Adelita V. Ranchor
Robbert Sanderman
Joke Fleer, and Maya J. Schroevers
Graduate School for Health Research
University of Groningen, the Netherlands

Background
The current study examined construct validity of a psychological scale. The authors are psychologists.

Aims
Several self-report questionnaires exist to measure mindfulness, with the 39-item Five Factor Mindfulness Questionnaire (FFMQ) being one of the most frequently used scales. The aim of the current study was to examine whether the scale items are understandable and whether the total score can be used to identify people with a different level of mindfulness.

Sample
The study was carried out in 152 undergraduate students.

Methods
Students were randomized in two conditions: 71 students filled out the original scale with five response categories and 81 students were given an extra response option (6 "never paid attention"). When students chose response option 6 at least once, they were asked to give reasons for choosing this option, and then to rate the items again, using the original 5-point rating scale. Rasch model was used to examine items' ability to discriminate different levels of mindfulness.

Results
Option 6 was indicated by more than 5 % of the students in 7 of the total 39 items. Rasch analysis showed items in four of the five subscales were less discriminative: the responses were either less accurate than expected or falling out of the predictable range.

Conclusions
The construct validity of the FFMQ is in doubt. Half of the items are either not understandable or not able to identify people with a different level of mindfulness.

No. PROMS2012PK001

Rasch Calibration of General Science Test at Grade-VIII in Pakistan

Presenter **August 7th (16:25–16:50)**

Muhammad Javid Qadir*, Abdul Hameed
School of Social Sciences and Humanities,
University of Management and Technology
Lahore, Pakistan
e-mail: javid_qadir123@yahoo.com
abdul.hameed@umt.edu.pk

Iram Gul Gilani,
Department of Education
Bahauddin Zakarya University Multan
Pakistan
e-mail: iramgul@bzu.edu.pk

The present study aimed at calibration of General Science achievement test for grade (VIII) through Rasch Model. For this purpose a General Science achievement test comprising 45 items was constructed from the text book of General Science for class VIII. Finally the test was administered to 300 students (M/F) in different high schools for boys and girls in Multan District. The answer sheets were scored and results were tabulated. Eleven (11) items were rejected on the basis of F, D and ϕ. Fifteen (15) items were to be improved on the basis of F, D and ϕ. Remaining all items were good items. Reliability value of the test was (0.82) and (0.85) by using Kuder Richardson # 20 and Kuder Richardson # 21 formula, which is very close to standardized value. Rasch Model indicates that overall test is good to measure the achievement of the students class (VIII) in the subject of General Science.

Keywords Calibration • General science • Achievement test • Rasch • Item analysis

PROMS2012RS001

Validation Study of Cut-Scores in School Achievements' Monitoring (SAM) Toolkit

Presenter
Наталья Гапонова
Center of International Cooperation in Education Measurement
e-mail: g.natalia999@gmail.com

Data collection during mass testing always leads to presentation of performance results. Establishment of cut-scores and division of examinees into groups due to their test performance is one of the most common ways to solve the problem of interpreting the results. However standard-setting procedure and its outcome are one of the most controversial which is difficult to validate empirically. Within each project that implements education assessment there is a professional and ethical responsibility for the results presented. That is why it is crucial to investigate the problem of how to provide examinees with valid and reasonable interpretation of test results. The first step in validation of benchmarks is development of standard-setting methodology for SAM, which results in establishment of cut-scores. The next step is to investigate internal, procedural and external validity of benchmarks. We investigated consistency between theoretical judgments and empirical estimates. Judges needed to set hypothetical p-values for each item as if examinees answered. Then we investigated the correlation between theoretically expected p-values and empirical within three different levels of competency. Secondly, validity evidence was investigated with the help of comparisons with other sources of evidence, such as consistency with results from, widely used method of standard-setting, Angoff. Thirdly, in order to prove procedural validity every step of standard-setting procedure is articulated clearly, including the purposes and processes. Moreover, the extent to which repeated applications of benchmarks through times get the same distribution of examinees on different levels was searched to prove internal validity. To conclude, we expect to prove the validity of the SAM's benchmarks, so that examinees could be sure that the interpretation of test results is rather valid, reasonable and precise.

No. PROMS2012SG001

Exploring the Features of Adaptive Neuro-Fuzzy Inference System and Path Analysis in Determining Factors Influencing Item Difficulty: A Study of While-Listening Performance Tests

Presenter **August 7th (14:15–14:40)**
Vahid Aryadoust
Centre for English Language Communication,
National University of Singapore,
10 Architecture Drive, Singapore 117511

Language researchers have adapted a variety of statistical tools to explore the variables that influence task difficulty in listening comprehension. Three notable examples are regression models, rule space methodology, and the fusion model. Although these studies have informed the field of language assessment, the focus of the majority is limited to post-listening performance (PLP) tests and their methodologies have limitations. PLP tests demand answering test items after the auditory experience. By contrast, a group of tests are while-listening performance (WLP) in which test takers are exposed to oral texts and engaged in simultaneous reading and answering.

This study reports a novel application of a class of neuro-fuzzy models (NFMs) called Adaptive Neuro-Fuzzy Inference System (ANFIS) into the International English Language Testing System (IELTS) listening test, a WLP test, and compares the findings with path analysis. NFMs are powerful artificial intelligence technologies which embrace the tenets of Artificial Neural Networks (ANN) and fuzzy set theory. The synergy of ANNs and fuzzy sets in NFMs provides promise for language assessment. In this study, seven variables influencing task difficulty were flagged during the item coding stage, a finding which was further supported by the ANFIS and partly by the path analysis. Results show that the NFM technique is promising in the context of language assessment. Further applications of the model in language assessment are discussed.

No. PROMS2012SG002

Synthesizing Language Measurement Advances

Presenter **August 6th (16:45–17:10)**
Vahid Aryadoust
Centre for English Language Communication
National University of Singapore
10 Architecture Drive, Singapore 117511
65–6601 2504(DID):: 65–6777 9152 (Fax):: elcsva@nus.edu.sg (E):: www.nus.edu.sg (W):: Company Registration No: 200604346E

This paper seeks to review the contribution of three major measurement trends to language assessment: classical testing theory, items response theory, and structural equation modelling. It is argued that classical testing theory and latent trait approach have propelled forward language (and educational) assessment since the 1970s. The paper further discusses the advantages of the latent trait models, specifically the Rasch model, differential item functioning, and cognitive diagnostic assessment models such as the fusion model and rule space methodology. Subsequently, it elaborates the construct modeling approach (Wilson, *Constructing measures*: *An item response modeling approach*. Mahwah: Erlbaum, 2005), as a useful method to develop language tests. The author discusses how Rasch-based construct modeling approach would assist language testers to develop second language listening comprehension tests. It finishes off by offering recommendations for future research in the field of language assessment and stressing the need to adapt novel measurement methods into language testing.

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

No. PROMS2012SG003

Evaluating the Quality of Ratings in Writing Assessment: Indices from Rasch Measurement Theory

Presenter **August 6th (15:30–15:55)**

Susan Tan and Chew Moh Leen
National University of Singapore

Stefanie A. Wind and George Engelhard, Jr.
Emory University

Background

The quality of ratings assigned in performance assessments must be systematically examined to ensure that ratings are valid, reliable, and fair for all students. Research on rating quality in performance assessment typically evaluates raters in terms of agreement, error, and systematic bias as indirect indicators of accuracy (Murphy and Cleveland 1991; Johnson, Penny, and Gordon 2009). Rasch Measurement Theory (RMT) provides a variety of indices for monitoring the quality of ratings in writing assessment, including indices of rater error and direct measures of rater accuracy (Engelhard 1994, 1996; Wind and Engelhard 2011).

Aim

The goal of this study is to examine the quality of ratings over two administrations of a placement test and to determine how such ratings affect student placement in a large-scale writing assessment.

Sample

This study uses essay scores from two English Placement Tests in Singapore. Fifty scripts were used from the 2010 test, scored by 37 operational raters and 30 scripts were used from the 2011 test scored by 32 operational raters. In addition, for each data set there is a team of three benchmark raters.

Methods

Differences in rating quality were considered in terms of rater experience, types of course taught and behavior over the two administrations. Many-Facet Rasch Models were used to examine raters in terms of error and accuracy. Indices of rater accuracy include rater severity calibrations and model-data fit statistics, direct measures of a match between operational and "benchmark" ratings. All analyses were performed using Facets (Linacre 2010).

Results

Preliminary analyses show differences in rater severity, model-data fit, and direct measures of accuracy which suggest that rater characteristics may influence the quality of student scores.

Conclusions

Differences in rating quality hold implications for research, theory, and practice. Inconsistent rating quality implies that the quality of inferences informed by rater-assigned scores may not be invariant across the raters.

PROMS2012SG004

Effects of Three Dimensional Visualizations on Learning Molecular Biology

Presenter **August 6th (14:45–15:10)**
Sandra TAN,
Hwa Chong Institution, Singapore
Russell WAUGH
Graduate School of Education
University of Western Australia
e-mail: sandra@hci.edu.sg

This paper describes the design considerations, implementation and results of formulating an approach to help students "see" DNA, proteins and cellular structures. An initial sample size of 249 Secondary three biology students was drawn from seven schools in Singapore. Purposive sampling was done in the selection of schools with different academic ability levels and yet fairly representative of biology students in Singapore. The experimental study leverages on novel immersive virtual reality technologies to help students understand the three-dimensional structures and the molecular interactions between them that enable function. Participants were taught topics in molecular biology either by traditional classroom "chalk-and-talk" or by a series of three-dimensional visualisation exercises using ICT. The students were tested for visual spatial ability and molecular biology achievement both before and after the intervention and the data analysed using a Rasch measurement model. Results indicate increases in visual spatial ability and molecular biology achievement. These effects were particularly pronounced in male students. Focus group interviews reveal that, prior to this intervention, students relied heavily on memorisation. This observation corroborates well with the analysis that revealed that students were well-trained in memorising specific biology phenomenon, but could not link these to general concepts and theory. The visualisation exercises helped clarify understanding while increasing interest and engagement. The results of this study recommend the use of technology in the teaching and learning of concepts in molecular biology, especially for male students in Singapore.

PROMS2012SA001

Rating Scale Model; Attitudes Toward Lesbians and Gay Men Scale; Dimensionality; Acquiescent Response Bias

Presenter **August 7th (13:00–13:25)**

Prof. Dr. Gideon P. de Bruin
University of Johannesburg
Box 524, Auckland Park, Johannesburg, South Africa
e-mail: deondb@uj.ac.za

Dr. Marlene Arndt
University of Johannesburg
Box 524, Auckland Park, Johannesburg, South Africa

Prof./Dr. Trevor G. Bond
James Cook University

Aims and Keywords
We examined measurement disturbances in the Attitudes toward Lesbians and Gay Men Scale (ATLGS).

Sample
Participants were 1,192 students (516 men and 671 women) who represented six faculties at a university in Johannesburg, South Africa. Participants' ages ranged from 17 to 42 years. Students represented four ethnic groups: Black, White, Mixed, and Asian.

Methods
Responses to the 20-item ATLG were fitted to the Rasch rating scale model. There are two subscales (Attitudes towards Lesbians and Attitudes towards Gay men). Items have a 9-point response scale.

Results
Four potentially problematic disturbances were detected: (a) a response scale for the items that calls for too fine a distinction in different attitudes, (b) a small but non-negligible group of persons who respond inconsistently across the Attitudes Toward Lesbians subscale and Attitudes Toward Gay Men subscale, respectively, (c) a small but non-negligible group of persons who respond inconsistently across favourable and unfavourable items, respectively, and (d) persons and items that do not fit the requirements of the Rasch model.

Conclusions

Better measurement with the ATLG can be achieved by using a response scale with fewer response categories. The use of favourable and unfavourable items introduces unwanted multidimensionality (possibly due to acquiescent response bias) and poor fit. Finally the results show that for most persons the ATLG measures a unidimensional attribute. However, for a small group of persons the total score is an invalid indicator of their attitudes. Rasch analysis can be used to identify these persons and other persons who respond in unexpected ways.

No. PROMS2012TW001

Intraclass Coefficient to Report the Strength of Dimension Using Rasch Principal Component Analysis on Standardized Residuals

Presenter
Tsair-Wei CHIEN
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw
http://facebook.com/smile.Chien.TW

August 6th (15:30–15:55)

Background

Dimensionality is an important assumption in item response theory (IRT). Principal component analysis (PCA) on standardized residuals has been used to check dimensionality, especially under the family of Rasch models. Smith (2002) proposed the percent of two sets of person ability with significant difference beyond 5 % that is deemed any potential multidimensionality occurred and can be supplemental to the traditional criteria such as a cutoff of 60 % of the variance explained by the Rasch factor, eigenvalues smaller than 3 and the percentage variance explained by the first contrast of less than 5 %. However, the message regarding dimension strength is required to report in a study.

Aims and Keywords

Following Smith (2002) rule on two sets of person ability, we proposed another way using intraclass coefficient (ICC) to verify dimension strength of a scale.

Keywords Intraclass coefficient • Rasch model • Unidimensionality • PCA • Cronbach's alpha

Sample

The Rasch rating scale model was used to analyze the 2009 English inpatient questionnaire data regarding patient satisfactory perception, which were collected from 162 hospitals, examined unidimensionality.

Methods

We developed a visual plot in Excel incorporated with WINSTEPS to automatically compute ICC, Smith's percent of two sets of person insignificant ability and its correlation coefficient that were compared with the results yielded by the responding simulation data.

Results

ICC could be as an indicator to measure the strength toward unidimensionality for a scale, through which 0.80 can be a cut off to discriminate whether a scale is unidimensional.

Conclusions

It is required to report the dimension strength of a scale when one factor is extracted from data that is like Cronbach's alpha reported as usual. In this study, we demonstrated dimensional validity for a scale with ICCs that might respond to the question about the standard format for reporting Rasch analysis in publications (Linacre 2010).

No. PROMS2012TW002

A Simple Screening Tool for Dengue Fever in Children

Presenter **August 6th (13:30–13:55)**
Wen-Pin Lai, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw
http://facebook.com/smile.Chien.TW

Background
Dengue fever (DF) is a significant public health issue in Asia. Some studies using univariate approach report that the presumptive diagnosis of DF is so imprecise because the signs and symptoms are not useful for detecting DF. Other multivariable regression analysis was attempted to distinguish patients with dengue from those with other fibrile illness; however, none had significant statistical validity and none considered changes in clinical features over the course of illness.

Aims and Keywords
We aimed to utilize clinical and laboratory data to derive a rapid and accurate DF case-finding tool for children.

Keywords Dengue fever • Dengue serologic test • Rasch analysis • Receiver operating characteristic (ROC) curve

Sample
This retrospective study used 24 DF-related characteristics and clinical features (17 clinical; 7 laboratory) from 177 pediatric patients (69 with DF diagnosis). Data were psychometrically evaluated and their effectiveness and accuracy in predicting DF risk were also evaluated.

Methods
Guided by the DF literature, a total of 24 DF-related clinical feature (7 laboratory and 17 clinical) capturing clinical, historical, and laboratory indicators were selected from patients clinically suspected of DV infection at the emergency department for the construction of a scale to screen for DV infection. According literature regarding symptoms of DF, data were obtained from the patients' medical records. Rasch principal component analysis (PCA) on standardized residuals were analyzed with WINSTEPS to search an acceptable combination of items to distinguish DF occurred.

Results
The 14-item scale (DF-14) was found to fit the measurement model in assessing DF tendency. The sensitivity (specificity) of the DF-14 measure was 0.759 (0.855) with a cut-off point of −1.15 (in logit), and the area under the ROC curve was

0.932 (95 % CI: 0.885–0.965). The DF-2 scale, comprised of white blood cell and platelet counts data, was brief but less comprehensive.

Conclusions
Simple laboratory data, such as those in the DF-2 and DF-14, can be useful for the early detection of DF risk in children. The DF-14 scale can be helpful in discriminating DF from other febrile illnesses before conducting a costly and time-consuming dengue confirmation test.

No. PROMS2012TW003

Rasch Analysis of the Relationship Between Physical Therapy and Occupational Therapy Manpower in Taiwan

Presenter
Hing-Man Wu, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw
http://facebook.com/smile.Chien.TW

August 6th (15:55–16:20)

Background

Manpower supply is an important issue in healthcare service industry. Whether the physical therapists in Taiwan are sufficient is required to study.

Aims and Keywords

The aim of this study is to investigate (1) the difference between physical and occupational therapy manpower in various medical settings and different insurance bureaus of geographical location, (2) the correlation coefficient between physical and occupational therapy manpower, (3) the performance for physical and occupational therapy manpower using a graphical representation in order to fulfill the requirement of the new version of 2011 hospital accreditation.

Keywords Physical therapy • Occupational therapy • Correlation coefficient • Graphical representations • structure equation modeling.

Sample

All the physical and occupational therapists registered in healthcare industry were recruited in this study.

Methods

Web-query system for medical practitioners in Taiwan is retrieved to obtain the numbers of physical therapist, occupational therapist, and hospital bed. Using the ratio of manpower supply to hospital beds, we examined whether it is adequate to meet the patient needs in hospital. The maximum of hospital capacity reaches to 2,918 beds and the minimum is 10 beds. The Rasch analysis with WINSTEPS and structure equation modeling are used to inspect the association between physical and occupational therapists in manpower supply.

Results

The results showed that the ratio of physical and occupational therapists in manpower supply was 2.54. The largest number of physical and occupational therapists serving for medical settings was in clinics (34.49 %), the least was in medical centers (12.64 %). A significant association was found in types of institutes related to rehabilitation professionals (Chi-square = 892.06; $p < .0001$), indicating interdependent with each other.

Conclusions

The data of medical practitioners and hospital beds retrieved from the website of Department of Health could be useful to compare the manpower of rehabilitation professionals using graphical representation. In addition, the method we introduced could be helpful for researchers to consider and explore in future studies.

No. PROMS2012TW004

Construction and Evaluation of an Item Bank for an Introductory Statistics Class: A Pilot Study

Presenter **August 6th (16:20–16:45)**

Sieh-Hwa Lin
National Taiwan Normal University
e-mail: linsh@ntnu.edu.tw

Pei-Jung Hsieh, (corresponding author)
National Academy for Educational Research
e-mail: peijung2@gmail.com

Background
Introductory Statistics is a required course for psychology and education related fields' students. However, many students are uneasy about the learning materials.

Aims
The purpose of the present study is to construct and evaluate the applicability of an item bank for an Introductory Statistics class.

Sample
The participants of the study were 54 college students enrolled in Statistics in Psychology and Education in the spring and fall semesters of 2010.

Methods
To establish the test bank, the authors first adopted and re-wrote the questions from the midterm and final tests of Statistics in Psychology and Education from the previous school year (2009). Students could practice the test in advance of the class on the instructional platform, Moodle. A total of 15 units were prepared with 45 questions. The collected data were analyzed by WINSTEPS 3.70.

Results
The results revealed that (1) the point-biserial correlation of 34 items (75.6 %) reached .25, meaning the test items constructed for this study had enough discriminatory power; (2) 80 % of the Rasch item difficulty values ranging from −1 to 1, indicating an appropriate difficulty of the test bank, which not only met the students level, but was also appropriate to help students with preparation for the unit.

Conclusions
Students' performance on the course score correlated positively with the number of times they took the preparation tests, suggesting that the implementation of the lesson preparation activity enhanced the learning effectiveness of statistics. In the future, the test bank will serve as the supplementary material for Introductory Statistics in psychology- and education-related fields.

No. PROMS2012TW005

The Aesthetic Intention Scale: Using the Health Belief Model to Predict Aesthetic Intention and Action Taking Among Hospital Employees

Presenter **August 7th (13:50–14:15)**
Shih-Bin Su, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw
http://facebook.com/smile.Chien.TW

Background
An aesthetic intention scale should be examined and applied to know the mostly possible potential customer would like to receive aesthetic treatment for a marketing plan.

Aims and Keywords
To develop an aesthetic intention scale which can predict aesthetic intention and action among hospital employees is incorporated with a report card disclosing individual examinees' aberrant responses on items.

Keywords Perceived susceptibility subscale • Improvement essentiality subscale • Rasch model • Item response theory • Aesthetic intention scale

Sample
The setting was a 900-bed hospital in southern Taiwan. A total of 1,800 full-time workers in the studied hospital participated in an aesthetic intention survey in May of 2009. The effective sample size was 1,124 with a *return rate of 62.64 %*.

Methods
Item response theory based Rasch model with WINSTEPS software was performed to examine an aesthetic intention scale whether forms a single construct. Multiply regress analyses were used to seek factors in association with aesthetic intention. Receiver Operating Characteristic (ROC) was applied to determine a cut-off point to looking for examinee's aesthetic action taking.

Results
The 12-item aesthetic intention scale has been detected to have two characteristic factors (i.e., subscales of perceived susceptibility and improvement essentiality). The cut-off point is set at 2.975 after combining those two scaling scores.

It accounts for 87.1 % of accuracy with sensitivity. The report card for investigating examinee's aberrant responses on items using standardized residual scores can provide institutes of aesthetic service with a diagnostic tool to inspect any response answered unusually in comparison to model's expectation.

Conclusions

The aesthetic intention scale examined by Rasch model can be used to predict aesthetic intention and action taking among hospital employees.

PROMS2012TW009

The Measurement of Creative Teaching Competency: Development of an IRT-Based Scale

Presenter **August 8th (9:25–9:50)**
Yu-Shu Chen
An associate professor at National Chung Cheng University in Taiwan ROC
e-mail: xiaoshu700@gmail.com

Yuan-Chi Lai
An assistant professor at Wu Feng University in Taiwan ROC
e-mail: yclai@wfu.edu.tw

Background and Purpose

Traditionally, Asian countries emphasize rote learning and passing examinations. The lack of creative teaching performance has had an adverse impact on human resource education. Teachers are required to produce competent graduates, ensure that the necessary technical knowledge is acquired so that graduates can effectively contribute to the workforce. How to improve human resource education hinges on teachers' creative teaching competency. However, there is no instrument to measure teachers' creative teaching competency. The major purpose of this study is to develop a reliable and valid scale to assess teachers' creative teaching competency.

Methods

The study was administered in two stages. In the first stage, based on Spencer and Spencer's (1993) competency developing model, the researchers developed a creative teaching competency scale which including seven dimensions: applying team learning, creating safe environment, deferring judgment, tolerance for ambiguity, accepting frustration, playfulness and humor, and risk-taking. The scale consists of 33 indexes. Then, the researchers converted the indexes into Likert-type items. For each of the 33 items, a response was sought in two perspective—importance and possession. In the second stage, we recruited 400 teachers from elementary schools and junior high schools. Two hundred of them are normal teachers and the other 200 are creative teaching award winners. Data were analyzed with a between-item multidimensional model because the scale contains seven dimensions and each measuring related but distinct latent dimensions. In this study, each item belongs to only one particular dimension, and there are no items in common across the dimension.

Results and Conclusions

The results yielded good psychometric property for the scale and revealed that the scale is appropriate for assessing teachers' creative teaching competency. The

seven-dimensional creative teaching competency scale has good reliability and discrimination. The scale showed substantial power for the explanation of variable in creative teaching performance. Educational implications of the current findings and suggestions for future studies are also addressed.

Keywords Creative teaching competency • Item response theory • Multidimensional between-item model • Rash model

PROMS2012TW010

Exploring and Modeling the Multidimensionality of Metalinguistic Knowledge

Presenter **August 7th (14:15–14:40)**
Wen-Ta Tseng, PhD
Associate Professor, English Department
National Taiwan Normal University
e-mail: wttseng@ntnu.edu.tw

Hsiao-Ling HSU, Doctoral Student, English Department
National Taiwan Normal University

Background

With many studies focusing on distinguishing between implicit and explicit knowledge, there is little research using the Rasch Modeling and the structural equation modeling to examine the explicit knowledge in particular.

Aims and Keywords

This study was aimed at examining the trait space of and confirming the structural relationships underlying the metalinguistic knowledge.

Keywords Metalinguistic knowledge • Rasch Modeling • CFA • MIMIC

Sample

Two hundred Taiwanese university students were recruited as the participants of the present study.

Methods

To measure metalinguistic knowledge, two types of test were adopted—untimed GJT (grammatical judgment test) and MKT (metacognitive knowledge test). Untimed GJT items were adapted from Loewen (2009) and MKT items were adopted from Lin (2009). Both GJT and MKT contain 20 items respectively, and each test was further sub-divided into three groups based on their feature of grammaticality— that of vocabulary, basic English grammar, and functional words. The trait space of metalinguistic knowledge was checked by the Rasch Modeling, and the structural relationships underlying it was examined by the Confirmatory Factor Analysis and MIMIC model analysis. Winsteps 3.72 was utilized to execute the Rasch Modeling, and AMOS 18 was then further implemented to perform a series of CFA model comparisons and MIMIC analysis.

Results

The results from the Rasch Modeling showed that the latent trait under investigation could be considered a bi-factor model. Based on this initial but fundamental finding,

the CFA analysis with this confirmed bi-factor model could also be well-supported by a number of model fit indices. More importantly, the results of the MIMIC model revealed that from a developmental perspective metacognitive knowledge has moderate predictive power over the functioning of learners' grammaticality judgment in English.

Conclusion

Both the theoretical and pedagogical implications were duly proposed in response to the research findings.

PROMS2012UK001

Is Analysis Under the 3PL Model of Practical Value to Test Makers?

Presenter **August 8th (9:50–10:15)**
Jeffrey Stewart
Kyushu Sangyo University, Cardiff University
e-mail: jeffjrstewart@gmail.com

Background
Although the Rasch model assumes minimal guessing, it is commonly applied to multiple-choice test formats, for which some argue the 3-parameter logistic (3PL) model is more appropriate. Theoretical debate aside, it is necessary to empirically test if analysis under the 3PL model can result in practical improvements of test forms.

Aims
To answer the following question: If a test which exhibits high person separation and an ideal TIF under the Rasch model is redesigned in accordance to implications of the 3PL model, does the revised form demonstrate higher reliability under the criteria of Rasch measurement?

Sample
Two samples each of approximately 1,500 first and second year English conversation students at a Japanese university.

Methods
Using a bank of 300 items, a 100-item test was designed which, under subsequent analysis in WINSTEPS, produced an ideal TIF under the Rasch model and had a person separation of 3.18. However, in a model comparison in the R package ltm, the 3PL model attained superior fit, and analysis under it indicated the test was somewhat difficult for the target population. A new 100-item test was designed to produce an optimal TIF under the 3PL model, and it was also examined under WINSTEPS in order to compare its Rasch reliability to the previous form. Although the TIF under the Rasch model was somewhat skewed, Person separation increased to 3.65, higher than under the Rasch-optimized test form.

Conclusion
Even if the goal of a test maker is to produce a test calculated under raw score, if multiple-choice formats are used, analysis under the 3PL model can result in tests of substantially higher person reliability, even under the criteria of Rasch measurement.

No. PROMS2012US001

Measuring Cumulative Learning of Energy Topics

Presenter **August 8th (9:00–9:25)**
Ou Lydia Liu
660 Rosedale Rd, Mailstop R7
Princeton, NJ, 08541, United States
e-mail: lliu@ets.org

Background

This research is about measuring cumulative understanding of energy topics among middle school students. Author is Dr. Ou Lydia Liu, Research Scientist at the Educational Testing Service at Princeton, New Jersey, United States.

Aims and Keywords

The aims of the research are to develop curricula and assessment to help advance the learning and assessment of energy topics among middle school students.

We ask three research questions: (a) How valid, equitable, and reliable are the knowledge integration energy items developed by the Cumulative Learning using Embedded Assessment Results project? (b) Do the items provide evidence for cumulative learning, cross-sectionally and longitudinally? And (c) What is the impact of unit learning on cumulative learning.

Keywords Science assessment • Energy topics • Rasch partial credit model

Sample

Participants are 4,160 middle school students from four schools in California in the United States. The sample consisted of 1,186 6**th** graders, 1,807 7**th** graders, and 1,167 8**th** graders. The sample included 2,108 males and 2,181 females. Among the participants, 3,087 speak English as their first language and 1,202 speak English as a second language.

Methods

Five energy curriculum units were designed and taught to the students. After learning the units, students took the end of the year assessment in both 2010 and 2011. The psychometric quality of the assessment items was evaluated using both classical testing theory and the Rasch partial credit model. In addition, the performance was compared among students who didn't receive the instruction, and students with different exposure to the instruction.

Results

The assessment items showed satisfactory psychometric quality. Students who had more exposure to the energy instruction made bigger progress on the assessment than students who had less exposure to the instruction.

Conclusions

The assessment designed to measure cumulative learning of energy topics demonstrated satisfactory reliability and validity. The energy curricula were effective in terms of cultivating students' cumulative understanding of energy.

No. PROMS2012US002

Construct Mapping, Theory Building, and Validity Testing: An Examination of the Student Moral Character Scale

Presenter
Jade Caines, Ph.D.
University of Pennsylvania
Graduate School of Education
3700 Walnut Street, Room 335
Philadelphia, PA 19104–6216
e-mail: jcaines@gse.upenn.edu

Background

I was an educator for over 10 years, teaching Kindergarten through university students. In 2011 I received my Ph.D. in Educational Research, Measurement and Evaluation. Currently, I am a postdoctoral fellow at the University of Pennsylvania.

Aims and Keywords

In recent years, there has been a greater focus on the development of student character and how it influences performance in the classroom (Davidson, Khmelkov, and Lickona 2010). Limited research, however, exists on the validity of instruments used to measure character in students. To address this gap in the literature, this study uses the Rasch Rating Scale model to investigate the convergence of item fit and theoretical expectations of a scale that measures Student Moral Character (SMC).

Sample

Data were collected on 239 middle-school students. The student demographic data are as follows: 92 males (39 %), 134 females (56 %), and 13 Unknown (5 %); 180 Black students (75 %), 40 Hispanic/Latino students (17 %), 1 Native American student, 1 White student, and 17 students "Unknown" (7 %); 163 sixth graders (68 %), 56 seventh graders (23 %), 9 eighth graders (4 %), and 11 students "Missing" (5 %). Also data were collected from four different public schools that range in size and geographical location.

Method

The Rasch rating scale model was used. WINSTEPS was used to produce item level statistics and Wright maps.

Results

Results suggest that the SMC scale does not follow many of the theoretical expectations. Although all of the SMC items show mean square error statistics (OUTFIT values) that fall within the acceptable, restrictive range, some item disordering exists.

Conclusions

First, the Rasch model provides a useful framework for examining item fit statistics and theoretical expectations. Second, scale revisions should be made before inferences from this scale can be considered valid. Finally, implications include theoretical advancement for the educational measurement and character education fields.

No. PROMS2012US003

Metaphor as a Basis for Unifying the Conception of Measurement Across Physics and Psychometrics

Presenter **August 7th (13:00–13:25)**

William P. Fisher, Jr.
University of California, Berkeley
e-mail: wfisher@berkeley.edu

A. Jackson Stenner
MetaMetrics, Inc.

Background

New things lacking concepts come into words and take their places in a language's network of significations via a process through which unclear images or figures not yet signified are identified and located relative to everything else. In physics as in psychology, that process is metaphor.

Aims and Keywords

Does metaphor calibrate a virtual measure of meaning? A study of the metaphor "love is a rose" explored this question.

Sample

Thirty-six residents (18 men, 18 women) of two locations in the US state of Illinois rated 68 entailments of the metaphor ("love is a plant," "love is beautiful," etc.) on a six-point agree-disagree rating scale.

Method

A theory of the construct implied by historical study of the metaphor's entailments was devised. The explanatory capacity of this theory was experimentally tested. Items representing the major theoretical aspects of the construct were separated onto three survey forms. Items tapping the same theoretical aspect of the construct should calibrate to the same locations, within a 95 % confidence interval. Data were fit to a polytomous Rasch model, and mean item locations were compared using ANOVA.

Results

Model fit was satisfactory, person measure separation reliability was 0.89, and item calibration reliability was 0.92. The three groups of items calibrated in the predicted order, with the low and medium, and medium and high, group means both differing by about a logit (respectively, t-statistics were 3.4 and 5.2, with 38 and 36 d.f. and $p \leq .002$). Average measures by sex and location were statistically identical, as were the average calibrations by form.

Conclusions

Measurement is a complex array of interdependent metaphors involving an invariant, portable unit defined via a mathematical law, predicted from theory to a useful degree of precision, and efficiently deployed in practical applications in a universally uniform metric traceable to a reference standard.

No. PROMS2012US004

Contrasting Physics and Psychometrics: Units, Laws, Theory, and Metrology

Presenter **August 8th (9:00–9:25)**

William P. Fisher, Jr.
University of California, Berkeley
e-mail: wfisher@berkeley.edu

A. Jackson Stenner
MetaMetrics, Inc.

Background

The ongoing cultural and economic successes of the physical sciences stand as a model to be emulated by all fields. Similarly rigorous sciences of psychological and social constructs require close attention to qualitative issues of unit definitions, lawful data regularities, predictive theories, and metrology's social networks.

Aims

Illustrated contrasts of the different ways psychometrics and physics approach these issues may be instructive.

Results

For instance, the identification and maintenance of a unit has long been of central importance in physics, but has become of interest in psychometrics only in recent decades. A second difference involves the definition of measures in terms of nonarbitrary lawful regularities; qualitative relationships between magnitudes of different attributes, such as mass, force, and acceleration, are embodied in physical measures but usually not in psychometric ones. A third difference involves theoretical understanding of the constructs measured, where explanatory power informs a capacity to accurately predict the causal trade-offs resulting from interventions on one attribute while holding another attribute constant. In psychometrics, predictive theories facilitating control over constructs have been obtained, and are of increasing interest, but are not commonly pursued. A fourth difference involves the convening of international standards groups responsible for maintaining traceability to theory-informed constant unit values essential to scientific and commercial activities globally. No such groups or traceability mechanisms exist in psychometrics, where most instruments measuring a given construct do so in their own unique units, and there is little awareness that traceability of this kind might be viable and valuable in human, social, and economic terms.

Significance

Issues concerning units, laws, theories, and metrological networks may represent opportunities for advancing the state of the art in psychometrics.

No. PROMS2012US005

Measuring Nurse's Attitude Toward Communication Between Nurses and Physicians Using Construct Modeling and Rasch Analysis

Presenter **August 6th (14:20–14:45)**
Mary P. Bourke RN, MSN, PhD
Indiana University School of Nursing Kokomo
Office Location KE 312
e-mail: mbourke@iuk.edu

Aims and Keywords
Understand how Construct Modeling is used to identify dimensions of a construct under study-a guide for instrument development. Synthesize how Rash Model diagnostics are used to evaluate the effectiveness of instruments developed for measurement.

Keywords Construct modeling • Rasch model

Sample
The population used for this study was a convenience sample of 89 RN's currently working at a regional hospital.

Method
A partnership was formed between a large regional hospital and Indiana University for the purpose of research. The purpose of the research was to ultimately improve patient outcomes by understanding the dimensions of physician/nurse communication as perceived by the nurse. The research team developed an instrument based on Construct Modeling, thus identifying dimensions of communication within the context of medical care. The identified dimensions were used as a guide in the creation of items for the instrument. Rasch diagnostics were performed using Winsteps software. Several iterations of the instrument were developed after consultation with peers and analysis of usability test results. A final version of the survey was distributed to RN's who volunteered to participate. IRB approval was obtained.

Results
Tool diagnostics were performed as follows: category frequencies and average measures, as well as, threshold estimates, probability curves, and category fit statistics. Rasch analysis provided detailed information about the psychometric properties of the instrument.

Conclusions

Validity was articulated within the context of the Rasch model. Person and item reliability indexes were clarified in relation to defining reliability of the instrument and interpretation of the data.