International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 1-7
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1001
www.arcjournals.org

From GITEST to RASCH-GZ - Inheritance and Development of Rasch-based Research in China

Zhang Quan

Professor Zhang Quan, Ph.D. in program of applied linguistics under Prof. Gui Shichun at Guangzhou Institute of Foreign languages, China (1989-1993), core member of the matriculation English test (MET) equating project (1990-1999), deputy chief examiner of college English Test (CET) Band 4 and 6 at Guangdong Provincial level; director of the language testing institute of Jiaxing University, China; senior visiting scholar at ETS (2002-2002), senior research scholar at UCLA under Prof. Lyle F. Bachman (2016-2018); China mainland representative of the Pacific Rim Objective Measurement Society (PROMS), PhD supervisor of City University of Macau, SAR, China (2013-2019) and of Deep Education Institute, Wisconsin, USA (2019-) and reviewer of several international journals. Prof. Zhang has been actively involved in research and application of language testing ever since 1986 and has translated and published monographs, edited books and articles.

Corresponding Author:** ***Zhang Quan, *principal of City Training Institute (CTI), Guangzhou, China (1996 -), Professor of Jiaxing University and PhD supervisor of Deep Education Institute, USA (2018 -)*
qzhang141@rasch-gz.com

Abstract: *Rasch model was first introduced into China in the 1980s by Prof. Gui Shichun, the famous Chinese linguist. It is Prof. Gui who successfully applied Rasch model to the ten-year (1989-1999) MET equating project with the strong support of National Education Examination Authority (NEEA) under the Ministry of Education, China. The equating project played a vital role in the implementation of standardized tests in China and was recognized by American and British peer experts. That was "Good Old Days" for the language testing community in China. More than 30 years have passed, and Prof. Gui Shichun passed away on April 5, 2017. In order to comfort the seniors and work harder, the author diachronically reviews the history of standardized testing practiced in China, and focuses on two aspects: the application of GITEST to the MET equating project and the introduction to the RASCH-GZ, the fully upgraded GITEST to reflect the inheritance and development of Rasch-based research so as to further promote the application of the Rasch model to language testing in China.*

Keywords: *standardized test; test equating;* GITEST, RASCH-*GZ, Rasch model*

1. A GLIMPSE OF THE 40 YEARS OF STANDARDIZED TESTING IN CHINA

In 1977, China resumed the entrance examination for higher education in all the provinces with inconsistent examination time and different examination formats. The total number of candidates reached 5.7 million. In 1978, the unified examination was implemented across the country, and the admission was based on the score.

In 1986, Professor Gui Shichun published Standardized testing: theory, principles and methods, which laid the solid foundation for the standardized testing from the perspective of Classic Testing Theory (CTT), and played an important role in the effective implementation and control of large-scale testing with high-stake in practice. At the same time, the publication by Prof. Gui also gave scientific answers to clarify various negative comments in the society that distorted the practice of standardized tests; in the same year, Chinese government launched the pilot study of standardized test item production, and multiple-choice (MC) question format became the main type of test items used, and the quality of the test item writing was basically stable.

In 1987, the NEEA was officially set up under the Ministry of Education, China. Starting from 1990 to 1999, with strong support by NEEA, a qualified test equating team led by Prof. Gui Shichun successfully implemented the ten-year MET equating project. No technical mistakes whatever occurred during the project years. Over the past 40 years, 108 million people have passed the entrance examination and entered universities (Li as cited in Yang, 2017, CPPCC website).

This paper discusses the inheritance and development of the Rasch model from two aspects: the application of GITEST in the MET equating project and the introduction to RASCH-GZ, the newly updated GITEST.

2. THE RESEARCH BACKGROUND OF MET EQUATING PROJECT IN CHINA

Since the implementation of the unified examination in 1978, admissions have been based on raw scores. It seems holding water, but in fact this is the crux of the problem. From the perspective of the language testing profession, two issues existed that were the most controversial at the time: First, for such a large-scale examination with high-stake, how to put under control the overall test item difficulty inherent in each year? Secondly, it is unscientific to select the "best students" based on the original score. To rephrase, it is not scientific to simply add up the original scores of several tests (i.e., the scores on the test takers' papers) because the difficulty level of each test paper is different (Gui, 2017). These two problems must be solved without delay; otherwise, the standardized tests cannot be carried out in China. From a more professional point of view, these problems are closely related to the difficulty of test items. So the calibration of test item difficulty and the implementation of test equating were put on the agenda.

In 1985, Guangdong Province took the lead in carrying out the pilot study for the reform of standardized test of the matriculation English test (MET). From 1988 on, the Rasch model was tried to solve the problem of the test score equating for MET in Guangdong Province, China, and positive progress was obtained. Data collected at several observed middle schools show that it is feasible and necessary to use GITEST, Rasch model-based system developed by ourselves to conduct the equating (Gui et al., 1993).

However, to ensure the safety, we used GITEST, BILOG, and PARSCALE to do the equating and compared the results at the same time. Later, we found that the results obtained from GITEST had a very high correlation with the data obtained by both BILOG and PARSCALE. Therefore, over the past few years, GITEST was used only for the project. In other words, GITEST has played a pivotal role in the ten-year MET equating project in China. Figure 1 below shows the difficulty curves of GITST, BILOG and PARSCALE based on the same data (1990-1999).

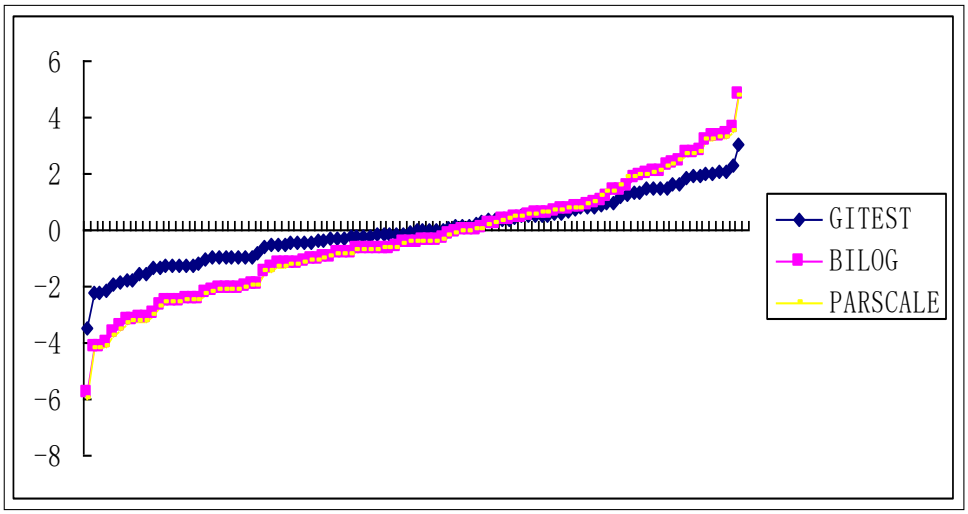

Figure1. *Difficulty curves of GITST, BILOG and PARSCALE based on the same MET data (1990-1999)*

As shown in the figure above, the three curves are very close. The two ones from BILOG and PARSCALE almost overlap. This has something to do with the number of iterations set in each command file and the preset value of convergence. BILOG came to convergence after 6 cycles of iteration with maximum change = 0.005; while PARSCALE converges after 72 cycles with maximum change = 0.01. The one from GITEST looks a little different. This is because all parameters of GITEST are set to default values. On the whole, there is no much difference in the calibration of test items.

From 1990 to 1999, with the strong support and leadership of NEEA under the Ministry of Education, the MET equating group led by Professor Gui Shichun successfully implemented the MET 10-year

equating project. This is the most difficult research project at China national-level, which is characterized by: the largest number of test takers in the world, equating based on the real data, admission based on the actual rescaled score reporting, long time span, nationwide coverage, and no errors, especially no technical error. The 10-year MET equating project was the equating model originally created by Professor Gui Shichun, which has been unanimously recognized by peer experts such as Charles Alderson and Lyle F. Bachman. What‟s more, the results verified the hyporthesis first proposed by Wright & Stone (1979) that the linking items are the “hard” items in the EASY test but the “easy” items in the HARD test. And the project enjoyed the reputation of “China’s Equating Model".

3. GITEST AND MET EQUATING PROJECT

GITEST (Gui, &Li, 1984) is the earliest Rasch-based software system developed in China. The use of GITEST laid a solid foundation for the successful application of the Rasch model to the test equating.

3.1. Equating Defined

The concept of „equating‟ discussed here refers to linking of test forms through common items so that scores derived from the tests which were administered separately to different test takers on different occasions after conversion will be comparable on the same scale. (Angoff, 1984; Hambleton & Swaminathan et al.,1985; Bachman, 1990; Kolen & Brenman, 1995, 2004; Gui,1990; Gui, Li & Zhang,1993, 2017; Li, 2000; Zhang, 2004).

3.2. Characteristics of MET

What characterizes MET in China can be illustrated in the five points as follows:

3.2.1. Compulsory exam: required for all Chinese high school graduates who plan to study at a university in China;

3.2.2. High risk: passing or failing to reach the cut-off score officially set determines whether a person can enter university or not;

3.2.3. Unified exam: the same test paper is used uniformly and administered in the same time period across the country;

3.2.4. English test paper is based mainly on MC question format plus a small part of writing;

3.2.5. From 1990 to 1999, the test equating was carried out three days once a year before the MET was administered, and the test results, after conversion within two weeks, can be compared on the same scale.

It is worth mentioning here that the situation in China is unique in at least the following three aspects (Gui, 1990):

(1) Due to the unbalanced development of education, there is a large number of test takers. Although they are all high school graduates, their overall quality is heterogeneous; therefore, it is difficult to set a fair (unbiased) test, let alone the equating of two sets of parallel test papers for different batches of candidates in different years.

(2) Although the test items of MET were centralized production, there is no way to afford centralized marking of the test paper according to the conditions at that time. The general practice was to assign each provincial examination authority to grade exam papers for candidates of its own province and to set its own admissions criteria. This has left universities faced with the problem of choosing admission scores based on different criteria set by different provincial examination departments.

(3) In China, there is no feasible way to ensure the safety of not exposing the contents of each large scale examination paper with high-risk after they were administered. Technically, the annual linking items cannot be changed, nor is it feasible to carry out any testing for future use. In order to find feasible yet safe solutions to these problems, sampling bases (middle schools) were established at that time to monitor the performance of fresh candidates.

Such a practice continued for 10 years (1990-1999). During those ten years, we can not only observe but also compare the performance of candidates who took the MET in different years. According to

Gui's (1990) assumption: within one year's time, there would be no big changes in terms of the general level of test takers. If there is any change, it must be associated with the change in the difficulty of the test items. Then we began to realize that such an assumption is by no means perfect, for at least three reasons:

First, there is the issue of sample size. We were going risk of test leaking. Statistically, the sample must be big enough to be representative; however, the larger the sample, the greater the risk of test leakage. Next, the overall level of candidates is unlikely to remain the same. Instead, it may fluctuate. Years of insignificant changes can be accumulated to significant changes. Finally, if there is any change in the difficulty of the test papers, it is unacceptable to make a linear adjustment for the differences between test papers only based on the candidates' individual test scores (Gui,1990).

3.3. Anchor Test Random Group Design

Given the above, we adopted the anchor-test-random-groups design. Specifically, the test results in 1988 (the first year when MET was administered in China) were used as basal test for calibration reference. Using the data from the sampling schools, test items of each subsequent test paper were equated to those of 1988 (i.e., item difficulty calibration and ability/score adjustment) (Wright, 1979). In layman's terms, if the test items were found to be more difficult than those in 1988. The relevant score would be increased; if found to be easier than those in 1988, the scores would be reduced.

3.4. Ability Estimation

In the case of the Rasch model, ability estimation is straightforward. To obtain a maximum likelihood estimate of theta (θ), we used the Newton-Raphson procedure (Hambleton, 1985). Ability values were again converted to probabilities for the general public who does not follow Rasch. Since the Rasch model has out-of-sample features, we can leverage the derived data to obtain an adjusted score for the population.

3.5. Rasch Model Preferable

Why was the Rasch model chosen over other models such as two- or three-parameter models of IRT? Its theoretical basis is as follows:

3.5.1. Feasible Implementation

Once the test items are calibrated, the relevant ability parameters can be estimated. Here's a typical example: Candidates who get a raw score for 60 correct answers out of 85 questions will be assigned an ability value regardless of the combination of those 60 correct answers. In contrast, in the case of two- or three-parameter models, the procedure becomes complicated. Estimation is closely related to discrimination and so-called "guessing" parameters. Therefore, since the combination of 60 correct answers varies from person to person, two or more candidates with a raw score of 60 correct answers out of the 85 test items will be assigned different ability values. Imagine, this would be a huge or astronomical number of items from 1 to 85! This makes it impossible to use sample data to predict the overall performance. And many items, the iteration never converged, mainly because of two big problems: the computer configuration problem at that time and the amount of data that could not be processed within two weeks‟ time (Gui, 1990). Even now, Professor Gui's approach is still scientifically acceptable. In 2021, PROMS[1] conference was held in Nanjing, China where Professor Steven Steiner, the keynote speaker[2], from the United States delivered the speech titled: "Better Measurement, Fewer Parameter! The true value of Rasch over IRT". The speech also confirmed Gui's approach.

3.5.2. Model/Data Fitting

Using GITEST based on the Rasch model, the item and ability fit can be estimated (Wright, 1982), which provides a strong demonstration of the goodness of fit of the Rasch model.

[1]Retrieved from http://proms.promsociety.org/2021

[2]"Better measurement and fewer parameters! The true value of Rasch over IRT", the keynote speech given by Professor Steven Steiner at PROMS2021, Nanjing, China.

3.6. Features of GITEST

Item analysis module used by GITEST is based on classic test theory. Each option of the MC question type was analyzed. Table 1 below shows the parameters of item analysis. And the equating module is based on Rasch model.

Table1. *Item analysis based on CCT and idea interpretation*

Item analysis	Ideas and interpretations
Mean	the mean scores of the whole examinees;
SD	the standard deviations of the whole examinees;
Varn	the variants based on the whole examinees;
P+	probability of correct answers;
Pd，	Δ value, difficulty parameter based on probability;
R11	by Kuder-Richardson20， reliability, this value should be over 0.9
aVALUE	reliability parameter, also called α value, by Cronbach formula, this value should be over 0.8
Rbis	discrimination index (in the unit of bi-serial)
Skewness	score distribution value, 0 indicating normal distribution; above 0, indicating positive skewness, showing the test items more difficult； below 0, indicating negative skewness, showing the test items easier；
Kurtosis	score distribution height: 0 indicating normal; above 0 showing “narrower”， i.e. small range between the scores; below 0, indicating “flat”, i.e. big range between the scores;
Difficulty	VD (<0.1), D (=0.1~0.3), I (0.3~0.7), E (0.7~0.9), VE (>0.9) VD: Very difficult； D: Difficult； I: Intermediate； E: Easy VE: Very easy

4. From GITEST to RASCH-GZ: Inheritance and Development of Language Testing in China

With the advent of the Internet era and the continuous improvement of computer technology and application requirements, the existing GITEST version can no longer meet the current needs. This is the motivation for us to comprehensively update and upgrade the GITEST system to RASCH-GZ during the global fight against the COVID-19 pandemic period. The Rasch model, powerful and feasible, will continue to serve language testing in China. To this end, since 2019, the author has organized a small yet qualified team from several universities and institutes to discuss, develop and update GITEST to meet the current needs and rapid development, so Rasch-GZ was born. The focus is to further promote the application of the Rasch model in the Chinese testing community so as to keep up with today's international practice. Meanwhile, it truly reflects the inheritance and development of language testing in China.

4.1. Comparison of GITEST with Rasch-GZ

GITEST and RASCH-GZ mainly focus on two major functions of language testing: item analysis and test equating. Table 2 shows the comparison between old version of GITEST and the fully upgraded RASCH-GZ system.

Table2. *Comparison of GITEST with RASCH-GZ*

GITEST	RASCH-GZ
BASIC，DOS	**(java, python, html, Delphi) online**
Data Matrix： 200 items by X 10,000 subjects (Maximum)	**Compatible with Excel: Unlimited items by unlimited subjects**
Key operating	**Menu operating**
Results in English，text file	**Results in both English and Chinese, WORD file**
Not applicable	**Plotting**
Not applicable	**Online technical support**

Figure 2 below shows the difficulty curve of the test items generated by both GITEST and RASCH-GZ after processing the same set of data respectively. The results show that the two curves based on the same data actually overlap 100%.

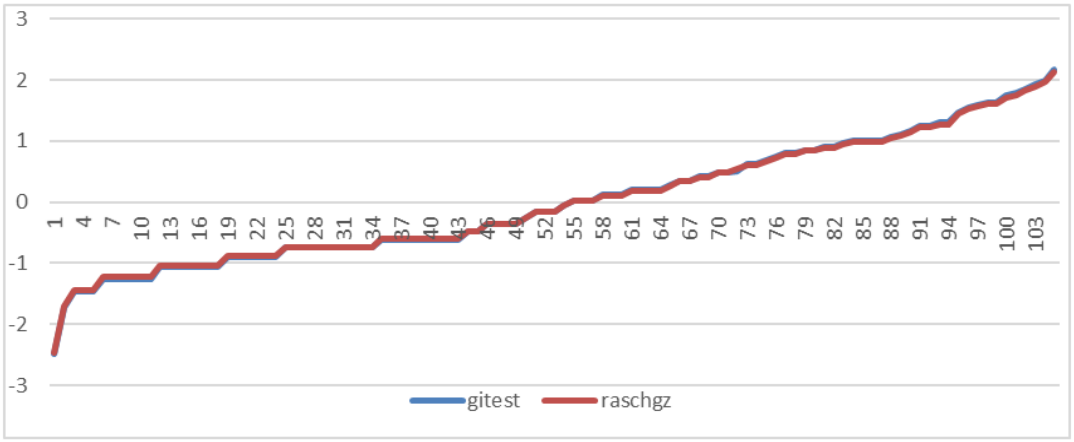

Figure2. *The difficulty curve of the test items generated by GITEST and Rasch-GZ on the same data processing*

In addition, according to the needs of practical applications for many years, some new functions have been added, which are mainly reflected in the following aspects;

(1) The interface design well meets the needs of non-English major researchers;

(2) After data file editing, the number of linking items can be flexibly selected;

(3) The system automatically performs the Chi-square test of linking items and deletes the items that do not meet the requirements;

(4) Data plotting function. The data can be plotted according to user's needs;

(5) Chinese and English language selection function. The data result files generated by the system can be selected to display or print in either Chinese or English, and

(6) Online technical support, etc.

5. SUMMARY

This paper briefly describes the occurrence and development process of the standardized test being practiced in China, focusing on the ten-year MET equating project, particularly the application of GITEST software and the introduction. To the fully updated Rasch-GZ in the recent years. From the perspective of language testing profession, GITEST, the earliest Rasch-based software system did play a vital role in the ten-year MET equating project. The application of GITEST and Rasch-GZ reflects the inheritance and development of the Rasch-based research in language testing in China. Rasch-GZ is the first Chinese version of Rasch-based item analysis and test equating system, which will greatly facilitate the popularization and promotion of the Rasch model learning and practice among Chinese scholars and researchers.

To conclude, let's quote Linacre (2016) that the Rasch model proposes a practical solution based on log odds transformation. But now many social scientists think it is too complicated, and many mathematical statisticians still think it is too simplistic. From the author's point of view, this phenomenon of the intersection and collision of literature and science has always existed in academia. However, the author should remind here that for scholars who study language testing or other liberal arts, being able to master the Rasch model to engage in their own research is already beyond the scope of pure liberal arts research. In the era of rapid development, in the context of scientific research in the era of big data, using Rasch model to process binary-valued data may not be more accurate, but it will definitely be more correct!!

REFERENCES

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

Fischer, G.H. & Molenaar, I. W. (1995). Rasch Models: foundations, recent developments, and applications. New York: Springer.

Frank, B. Baker (2001). The Basics of Item Response Theory. University of Wisconsin, ERIC. Clearing house on Gui, S.C. (1990). Notes on Itembanking (2).Modern Foreign Languages. Issue 4, pp.66-72.

Gui, S.C., Li, W. & Zhang, Q. (1993). The Application of IRT to test equating for MET in China. PP.391-393. In NEEA (Ed.). the 4th China Education and Examination Conference proceeding China Peace Press. ISBN 7-80101-089-2/G.64

Kolen, M. J., & Brenman, R.L. (1995). Test equating: methods and practices. Springer Vertag New York, Inc.

Li, W. (2000).MET in China: Reform, Explore and Practice. Higher Education Press.

Linacre, J. M. (2016). A user's guide to WINSTEPS Rasch-model computer programs: program manual 3.92.0. ISBN 0-941938-03-4.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Trevor, G. Bond & Christine, M. Fox (2015). Applying the Rasch Model: Fundamental Measurement in the *Human Science* (3rded). Routledge.

Trevor, G. Bond & Christine M. Fox (2018). Applying the Rasch Model: Fundamental Measurement in the *Human Science* (4th ed). Routledge.

Wright, B. D. & Douglas, G. (1975). Best test design and self-tailored testing. Research Memorandum.

Wright, B. D. & Stone, M. H. (1979). Best test design: Rasch measurement. MESA Press.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? Rasch Measurement Transactions, 6(1), 196-200

Zhang, Q (2019). Rasch Model : Research and Practice in China. In Myint, Swe Khine (Ed.). (2019). *International trends in educational assessment.* Brill｜Sense. Retrieved from http://catalog.loc.gov.

Zhang, Q (2012). Towards International Practice of Language Testing in China. Keynote speech given at the PROMS2012, Jiaxing, China, August 6-9, 2012.

Zhang, Q. (2011).Towards Better Interaction between Testing and Teaching. Keynote speech given at the 5th National TEFL/1st Mongolia TESOL Conference, Ulaanbaatar, Mongolia, Oct 7-9, 2011.

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

Citation: *Zhang Quan. "From GITEST to RASCH-GZ - Inheritance and Development of Rasch-based Research in China" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 1-7. DOI: https://doi.org/10.20431/2349-0381.09S1001.*

Copyright: *© 2022 Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*