




RASCH-GZ: The Most Updated Rasch-Based Research Development in China

Quan Zhang^{1,2} 

¹Jiaxing University, Zhejiang, China

²The World Sports University, Macau, SAR, China
qzhang141@aliyun.com; qzhang141@thewsu.org

Abstract: The present article introduced (RASCH-GZ) from the perspective of both Classical Test Theory (CTT) and Rasch Model. (RASCH-GZ) is the recently developed software based on Rasch model particularly used for item analysis and test equating. The system was successfully developed during the global fighting against COVID-19 pandemic period from 2019-2023. This updated professional software (also a platform) provides Chinese researchers with lots of help in the application of Rasch model, hopefully contributing to the development and popularity in Rasch-based research for language testing in China.

Key words: Rasch Model, Test Equating, Linking Items, Item Analysis, (Rasch-GZ)

1. Introduction

This article is based on my personal reflection on Rasch- and IRT-based computer software for language testing. It's a long story. That starts with my first visit to Educational Testing Service (ETS) in Princeton, NJ, USA in 2002. At that time, I was the first Chinese visiting scholar to ETS which was seen, in the eyes of ordinary Chinese students, as a merely educational testing company, but only professionals engaged in language testing know that it is actually a gathering place for the world elites of language testing and statisticians who set test items to test the world. It is during that period of time that I met, at T-building in ETS, several famous experts such as Professor Bob Mislevy, the (BILOG) program writer, Prof. Paul Holland, the expert of test equating and etc. Frankly speaking, at that time, few professionals across China knew what (BILOG) was. In my impression, academically (BILOG) indicates rigorous and technically offers better solution than other software I used. That was 22 years ago.

After returning to China, I continued to use (BILOG) and (PARSCALE) and also published some articles and books about the research using both the software [1],[2],[3],[4],[5],[6],[7],[8],[9], [10],[11],[12]. Over the past years, whenever using the software, I got an obvious feeling that though effective and dominant in objective measurement, software like (BILOG) could not be very well promoted in China, especially among students of English major or humanities. Specifically, users need to write command files, and the data file format should be arranged according to the data input format specified in the command file. If anything in the command file was written incorrectly, the system will display an error code, but these are all English prompts, and the result files are also entirely in English. In short, in terms of command file programming, though simple and easy for computer program writers, Chinese students of non-English major or of humanities find it difficult to fulfill the task because they need to learn programming in simple Fortran; on the other hand, in terms of interpreting results, students of science are not very proficient. Even if the result files generated by (GITEST)¹ software we developed were all in English, the users of which, as a matter of fact, were totally confined within a small band of Chinese professionals of language testing. Thus, the overall level regarding the application of language testing theory in China will not be effectively improved.

How time flies! I have been busy teaching, dealing with office work, and running around as the dean of now military, now civilian universities. I could not afford any time to consider my original intentions, fortunately till the global outbreak of COVID-19 in 2019, which gave me a chance to take a break. I sat down, thinking quietly: In 1969, Wright & Panchapakesan wrote (BICAL)², the first computer program based on Rasch Model. The making of (BICAL) offered two great contributions: (1) making the IRT application a reality and avoiding the embarrassment that IRT model was confined within theoretical talking among testing professionals due to the complexity of its mathematical algorithm in practice; (2) promoting the practical use of Rasch model in the US. Then the Chinese version of (SPSS) flashed through my mind, Then the idea occurred to me. Well, why not to have the first Chinese version of Rasch model or something? It is this original intention that greatly motivates me to completely update our old version of (GITEST) or to develop (Rasch-GZ). No sooner said than done. Since COVID-19 made us go nowhere and could only stay indoors, I organized either online or via wechat a small yet qualified team of computer engineers and language testing experts to fulfill my original intention. It took us more than good three years (2019-2023) to have developed the current (RASCH-GZ) for Chinese-speaking users. With (Rasch-GZ), we provide Chinese researchers and students with a good platform for CTT-based item analysis and Rasch-based test equating. In fact, we break the language barriers through such software.

2. Methods³

¹ For details of (GITEST) , interested readers may refer to of the present proceeding.

² The (BICAL) program was developed by Dr. Benjamin D. Wright and his students, with the first version appearing in 1969. This program uses Birnbaum paradigm to analyze tests under Rasch model. (BICAL) is well documented from the point of view of both underlying theory and its use for test analysis (See Wright and Sone,1979).

Technically, (RASCH-GZ) [13],[14],[15],[16] [17] has two functions: Item analysis and test equating. The present article introduces both parts with typical examples for each. (RASCH-GZ) used Delphi, JAVA and Python. Delphi has played its advantages in developing Windows desktop application system, which has realized the interface for visual data editing fully compatible with Excel, thus, making easier data editing in Excel and data importing into the system for analysis. At the same time, desktop applications have excellent support of Chinese language. The purpose is to provide visualization, networking, both Chinese and English language support and the Window system suitable for Chinese users. Python has its unique advantages in data processing, and JAVA in networking, which is stable and reliable. The item analysis module is based on the Classic Testing Theory (CTT). This includes multi-dimensional analysis: item difficulty and discrimination index (DI) of each option, etc. Test equating module is on Rasch model.

2.1. Item Analyses

Procedures for test item analysis have been developed through CTT [18] [19] [20] [21] [22]. The detailed item analysis report generated by (Rasch-GZ) can be referred to the user guide to RASCH-GZ⁴. This includes the calibration of item difficulty, DI and distracter analysis (DA) for multiple choice (MC) question type. To keep the scope of the article manageable, in what follows, only basic concepts will be addressed with examples.

Item difficulty. The item difficulty calibrated by Rasch-GZ falls into two types: The one based on CTT, i.e. the proportion of test takers who answer the item correctly (also called the ‘p-value’) [19] (p.76). Such a p-value serves as good reference for item moderating; the other based on Rasch Model.

$$P_i = \frac{\exp(\theta - l)}{1 + \exp(\theta - l)} \tag{1}$$

where the probability (P) for test taker to elicit a correct response to item (in the dichotomous case of ‘0’, meaning incorrect; 1, correct) is a function of the difference between the measure of a test taker’s ‘ability’ and an item’s ‘difficulty’($\theta - l$) [19] [20]. To quote Rasch, the Rasch model is a latent trait model. A person’s response to an item is determined by the difference between person ability and item difficulty measured on the same continuous scale. The unit used is in logit which makes equating possible.

Δ value. (RASCH-GZ) adopts the Δ value first used by Educational Testing Service (ETS) [21] [22].

$$\Delta = 13+4Z \tag{2}$$

³ The author will not be introducing the basic operation of RASCH-GZ system, for interested readers please refer to <http://www.rasch-gz.com>

⁴. <http://www.rasch-gz.com>

where Constant 13 refers to the zero of Z value, indicating 50% of the total test takers got the correct answers; Constant 4 refers to 1 of Z value, i.e. out of a normal distribution, 84.13% got the correct answers. This is arbitrarily decided to avoid any negative value and to ensure the parameter an integer. In this sense, if 99.87% of test takers got the correct answers, the Δ value would fall in the -3 position of the normal distribution, therefore we obtain:

$$\Delta = 13 + (4) * (-3) = 1 \quad (3)$$

indicating a very easy item; on the contrary, if 0.13% of test takers got the correct answers, the Δ value would fall in the +3 position of the normal distribution, thus we have

$$\Delta = 13 + (4) * (+3) = 25 \quad (4)$$

showing a very difficult item. Δ values vary from 1 to 25. **The greater the value, the harder the item.** For example, the Δ value of the first item being 14.76 in Table 1 below tells us the position of the item in the whole test paper, merely a little above the average.

Item discrimination. Item discrimination describes the relationship between the test takers' answers to the test item and their answers on the total test [19] [20] [21] [22]. In other words, discrimination index (DI) answers such a question: if test takers are scoring high on the test overall, are they also answering this item correctly? [19], (p.77). To rephrase specifically, their correct answer to this test item reinforces the DI; otherwise, weakens the DI. (RASCH-GZ) adopts the following formula to obtain DI:

$$r_{bis} = \frac{M_r - M_w}{S_t} \times \frac{P(1)}{y} \quad (5)$$

wherein

M_r = mean of the total of test takers who got the correct answers;

M_w = mean of the total of test takers who got the incorrect answers;

P = the ratio of the number of the test takers who got the correct answers and the total number of test takers;

y = in normal distribution, the vertical line that separates P from $(1-P)$;

S_t = the SD of the total score of all the test takers.

Interpretations for DI. There are different ways to interpret discriminative information. Popham (as cited in [19], p.77) suggested that a discrimination of 0.4 and higher is a sign of a 'very good item, and that anything less than 0.2 is a sign of a 'poor' item. In (Rasch-GZ), we take 0.3 -0.7 as acceptable items.

(Rasch-GZ) adopts biserial correlation coefficient. However, there are at least two limitations [20]: One is that such a value does not have any standard error of estimate, nor is it limited to a range within -1 to +1 like r (p.214) [20]. To ensure better

understanding, (Rasch-GZ) truncates values greater than +1 as +1 and assigns them the value of +1 and truncates values less than -1 as -1, and assigns them the value of -1 so as to keep the value range within -1 to +1 like r.

For correct answers (Keys), Less than 0.3 (Weak, and needs professional moderation or abandon it); Equal to or greater than 0.3 and less than 0.7 (Acceptable); Greater than 0.7 (Very good but difficult to achieve, which requires professional item writing).

For distracter options. equal to or less than 0.1 (Weak and needs professional moderation. Any distracter must produce some distracting effect);

Greater than -0.3 and less than 0.2 (Acceptable);

Greater than 0.3 (Too strong and needs moderation);

Distracter = 0 (Poor. No distracting effect. Moderate or discard it because zeroed distracter causes the difficulty of MC questions to decrease.

There are two more points that need to be clarified here: Negative DI indicates that the test takers scoring high on the test overall answered the item incorrectly, while the test takers with low scoring got the correct answers. It is imperative to check up the original test item. This is often due to the poor wording in the test item stem or the other options, etc.); DI = 1, which is the ideal discrimination! This indicates that all the test takers scoring high in the test answered the item correctly. In contrast, all the low scoring ones answer the item incorrectly. This can only be the explanation as the concept of DI in classroom teaching. In practical tests on large scale yet with high stakes, it is infeasible to obtain such a test item with DI being 1 for the correct answer.

The relationship between DI and difficulty. In practice, the manipulation of the relationship regarding DI and difficulty is of great importance. When using RASCH-GZ for item analysis, priority is given to DI. In other words, to use or to discard a test item largely depends on its DI. For example, a test item with 2.986 (logit), according to the Rasch model, should be treated as a difficult item. However, if its DI is very low (< 0.2). In such a case, it is generally recommended to discard the item, or to moderate it before going through another pre-test.

Distracter analysis. Apart from item difficulty and DI, (Rasch-GZ) also adopts the technique of 'distract analysis' to further analyze the quality of a MC question format. In language testing, by 'distracter' is referred to as the incorrect options designed by test item writers to distract test takers from the correct answer, or rather from the key. Item analyses of (Rasch-GZ) based on CTT analyze each option of MC questions and generate a test report based on the parameters (See Table 3). This provides a detailed report regarding each option of a MC question and analyses for the whole test paper based on the data collected from pre-tests so that teachers, test item writers, testing practitioners etc. could use the information to do moderation of the test item production, to adjust the item difficulties, or to simply decide whether to abandon those inappropriately designed test items.

For better illustration, we focus on the interpretation for item analysis in the first file: MFR Data 001.ia1 generated from (Rasch-GZ). Table 1 below shows the item analysis report for the first test item.

Table 1. MFR Data 001.ia1

Test Name	Part	Test D	Part D	Item D	R/N	A-Ser	B-ser	C-Ser	D-Ser	O-Ser
GD08	LTN 1	15.6 6	15.59	14.76	0.33	11.30	12.16	14.7 0	12.7 5	
Total	Key	A-D	B-D	C-D	D-D	A-N	B-N	C-N	D-N	O-N
100	C	- 0.29	-0.16	0.39	- 0.05	18	23	33	26	0

Technically, if the users entered 85 test items, (Rasch-GZ) would generate 85 tables for each item like Table 1 for detailed analysis. They include item difficulty and DI of all the options of a test item. Table 1 above is for the first item of listening part of a test administered to 100 test takers. As can be seen from the first row, we have, in order

- (1) Test name: GD08, (defined by the user);
- (2) Part, LTN1, referring to the first item in the listening part; ,
- (3) Test D, referring to the overall difficulty of the test, which is 15.66;
- (4) Part D, referring to the difficulty of the listening part, which is 15.59;
- (5) Item D, referring to the difficulty of an individual item (here is the first item of listening part) which is 14.76. (RASCH-GZ) adopts the Δ value (as seen in 2.1.2).
- (6) R/N refers to the ratio of correct answer (which is 0.33), showing not a very difficult item;
- (7) A-Ser refers to the score of the test takers who chose Option A.
- (8) B-Ser refers to the score of the test takers who chose Option B.
- (9) C-Ser refers to the score of the test takers who chose Option C.
- (10) D-Ser refers to the score of the test takers who chose Option D.
- (11) O-Ser refers to the score of the test takers who did not take any option of this item.

Since the key is C, the score of those who chose Option C as the correct answer is 14.7, hence a little bit higher than those who took other options.

Starting from the second row, we have, in order,

- (1) Total, referring to the total number of the test takers (here we have 100);
- (2) Key, the correct answer of the item (C);
- (3) A-D, refers to the DI of Option A;
- (4) B-D, refers to the DI of Option B;
- (5) C-D, refers to the DI of Option C;
- (6) D-D, refers to the DI of Option D;
- (7) A-N, refers to the number of the test takers who chose A (18);
- (8) B-N, refers to the number of the test takers who chose B (23);
- (9) C-N, refers to the number of the test takers who chose C (33);
- (10) D-N, refers to the number of the test takers who chose D (26);
- (11) O-N, refers to the number of the test takers who did not take any option of

the item (0).

Now, let's examine the four options A, B, C, and D of the test item. They are: -0.29, -0.16, 0.39, and -0.05, respectively, which, EXCEPT C, the key, basically do not meet our test requirements. Our interpretation here is that Option A and B are not well designed because they produced negative effect, showing that test takers who got higher scores turned out to take these two choices. Therefore, our test designer and test item writers are supposed to pay attention to such phenomena. And Option D is too weak. It did not act effectively as a distracter, therefore, it needs moderating. In the rigorous practice of item moderating, such an item should be deleted.

Based on the above, the author should say that -ial file is very important for professionals to use as a reference to moderate test items after a pre-test was conducted. Item analysis serves as good guideline for experts of language testing to decide whether a test item is to be put into item bank, needs moderating or to be deleted. In this way, so long as the collected data are true and reliable, and the subjects used for pre-test are homogenous, testing experts will take the data as the objective standard in the process of moderating the test items. In doing so, the efficiency will go higher, and the pointless subjective-oriented issue(s) will be reduced to null!

2.2. Test Equating

To keep the scope of the section manageable, the author will be first addressing briefly test equating and its concept and then present a simplified example to illustrate the specific procedures of test equating performed by (Rasch-GZ).

Test equating defined. The test equating realized via (Rasch-GZ) refers to such a practice: Linking of parallel 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, can be comparable on the same Rasch scale [21] [23] [24] [25] [26] [27]. The following arrangement indicates the idea:

Test takers of Group X take Test X consisting of L items with n items used as linking items;

Test takers of Group Y take Test Y consisting of L items with n items used as linking items.

In language testing, this is known as two parallel test forms thus designed, each with "n" anchor items and are administered to two different groups of samples drawn from the same population at either the same or different time. What is intended to achieve is to equate the metric of all the L items of the two tests and put them on the same scale [21] [1] [2]. To accomplish this, we use Test A as the basal test calibration and choose, from this basal test, n items ($n < L$) as linking items and put these linking items in Test B. The following array shows the idea wherein Item 27 through Item 42 in both tests are used as linking items. Totally, we have 16 items in each test. The following integer arrangement indicates the data entry structure of (Rasch GZ).

Test A 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42

Test B **27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42** 43 44 45 46 47 48
 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67

This is considered as the typical examples in terms of “equating of parallel tests”. As equating is a complicated process requiring enormous data processing, and manual calculation is by no means feasible, (Rasch-GZ) now offers an effective tool. In what follows, the author presents a pair of representative yet real data to demonstrate the complete procedure of how equating is complemented using (Rasch-GZ).

Test equating via (Rasch-GZ): A simplified example

So long as the data file is imported successfully, the user simply follow the menu by a single click of mouse, (RASCH-GZ) can complete the equating with Test X and Test Y. The results are listed in Table 2 below.

Table 2. GITEST: Linking Item Difficulties in logit of Test X and Y

ITEM	Test X	Test Y
0001	0.335	0.055
0002	-0.237	-0.978
0003	-0.073	-0.669
0004	0.154	0.118
0005	-0.018	0.118
0006	0.154	0.736
0007	-0.073	-0.429
0008	-0.237	-0.068
0009	-0.981	-1.174
0010	1.156	1.472
0011	-0.073	-0.852
0012	-0.028	-0.608
0013	0.462	0.311
0014	0.213	-0.068
0015	-0.449	-0.189
0016	-0.555	-0.669
M	-0.016	-0.181

Here we are interested in the means of the 16 linking items in the two tests. As observed in the above table, the two means of the same linking items turned out to be

different, i.e. -0.016 (logits) in Test X and -0.181 (logits) in Test Y. Now the question is: Why are the difficulties of the same 16 items different? Our interpretation is that the two test items to which these common items are connected respectively in Test X and Y are different. If $-0.016 - (-0.181)$, the difference obtained from the Mean1 minus the Mean2 is 0.165 logit, indicating the test items in Test X are a little bit easier than those in Test Y. That is why the means of the 16 linking items in Test A turn out to be more difficult than those in Test B. In other words, test items in Test X are 0.165 easier in logit than those in Test Y. “In such an example, the linking items are the hard items in **EASY** test but the easy items in the **HARD** test” [28] [29] [30] [21] [1] [2]. In this way, the difficulties of the other items in both Test X and Test Y got equated and can be comparable on the same scale as listed in Table 3 below⁵.

Table 3. Equated Item Difficulties

ITE M	Test A	Test B
0017	0.528	0.378
0018	0.273	0.661
0019	0.528	-0.369
0020	0.596	0.896
0021	-0.29	-0.548
0022	0.596	-0.669
0023	-0.237	-0.791
0024	-0.449	0.98
0025	0.667	0.118
0026	-0.073	1.258
0027	-1.445	-0.488
0028	-0.927	-0.309
0029	0.213	-0.309
0030	-0.29	0.055
0031	0.596	0.516
0032	0.596	-0.309
0033	-0.018	-0.488
0034	-0.344	1.068
0035	0.335	0.118
0036	0.154	0.661

⁵As for the specific operation of test equating using Rasch-GZ, interested readers may refer to the user guide to Rasch-GZ at <http://www.rasch-gz.com>

0037	-0.555	0.055
0038	-0.073	-0.852
0039	0.895	-0.791
0040	0.096	1.068
0041	-1.092	0.98
0042	0.977	

Unit: logit

3. Discussion and limitations

Today, in language testing practice, equating is considered as the prerequisite condition for computerized adaptive testing (CAT), item banking and for online testing in the Intern-based testing as well. Through equating, the changes of item difficulties in the test forms can be observed and equated, and the corresponding ability estimates across different occasions are thus re-scaled. The most significance inherent in test equating is the maintenance of security and fairness.

However, even though test developers attempt to construct test forms that are as similar as possible in both content and statistical specifications, the forms always differ somewhat in difficulty. According to [18] [19], the comparability of tests scores (or ability estimates) across different tests measuring the same ability is an issue of considerable importance to test developers, measurement professionals, and test takers alike. In China, let's take Matriculation English Test (MET) for example, MET is the most prestigious, competitive and large-scale examination of high stakes administered annually to approximately 10 million candidates across China. Its item difficulties and test security must be put well under control. If the same MET paper is administered repeatedly to different candidates nationwide annually to admit students for university studies in China, there is no way of protecting test security and fairness immediately after its first administration. On the other hand, it would be infeasible to administer two separate tests at once to the same group of candidates for the purpose of comparing the item difficulties (i.e, equating with same subjects). In this sense, equating plays a central role.

Another significance contributed by {Rasch-GZ} to language testing is that the test equating results justify the assumption first proposed by [29] that the linking items are the hard items in **EASY** test but the easy items in the **HARD** test".

(Rasch-GZ) needs more academic promotion in and outside China. Ever since its making, efforts have been made to run workshops or do presentations at international symposiums⁶. Some PhD programs⁷ have begun to attach importance to it as well. To

⁶(Going to do) PROMS 2024, August 19-20th Kuala Lumpur, Malaysia;

International Conference on Educational Assessment and Testing in the Age of AI, Dec.17-18th, 2023, Jiangxi, China;

PROMS 2023, August 28-30th, Macau, SAR, China;

The International Education Colloquium 2022, Miri, Malaysia;

PROMS 2021, online December, Nanjing, China;

do a better job in the days to come, we need more international exchanges and cooperations.

While the authors have achieved some preliminary and positive results, limitations remain; for example, comparisons of the features/analyses between Rasch-GZ and other Rasch-based software should be conducted. This needs data. Therefore, the relevant research will be addressed in separate paper.

4. Conclusion

In the current era of big data, AI and computer technology development, true testing is based on real and objective data, rather than relying on classroom teaching. Although modern test theory like item response theory (IRT) or Rasch model are well accepted by Chinese professionals of language testing, it acts mainly at seminars or in classrooms. The actual situation in terms of application is by no means optimistic. Bond and Zi Yan in [31] undertook a simple literature search in the China National Knowledge Infrastructure⁸ (CNKI) with “Rasch Model” as the key word. The number of Chinese publications with citations of “Rasch Model” across different disciplines between 1985 and 2016 is very small. According to [31], in contrast to the number of publication in the social science each year in China which is around 1.1 million, the number of researchers using Rasch Model (in language testing) remains like ‘a drop in the ocean’.

In terms of the application of computer technology to language testing, China's current situation is fully in line with the international practice; however, the testing theory and application remain largely at the level of CTT with linear or descriptive statistics in most schools and universities across the country. In this sense, both IRT and Rasch measurement need further dissemination and promotion. Therefore, the author wants to emphasize that what we are doing may not be necessarily more precious but must be more correct! So long as the method is correct, testing data objective in nature using Rasch model is of primary importance!

To conclude, Rasch model can provide a good solution to many problems of objective measurement encountered in the social sciences, and is appropriate for researchers in professional fields such as language testing. Thus (RASCH-GZ) has

The 6th International Conference on Language Testing and Assessment, Shanghai, China, 2021;

Symposium held on campus of Guangzhou Institute of Foreign Languages (Now Guangdong University of Foreign Studies) for the 90th Anniversary for Prof. Gui Shichun. Dec.19th, 2020, Guangzhou, China

PROMS 2013, August 3-5th, Kaohsiung, China

⁷ The PhD program of Applied Linguistics, Xi'an Jiaotong University, China;

The PhD program of Education, City University of Macau, Macau, SAR, China;

The PhD program of Education, Nueva Ecija University of Science and Technology (NEUST), the Philippines

The PhD program of Education, Tarlac State University (TSU), the Philippines

⁸<http://www.cnki.net>

greatly promoted the use of Rasch model among Chinese speaking researchers. The author hereby reminds our readers that the easiest way to learn how to use Rasch model measurement is to download the student version of (RASCH-GZ) and user manual from the

<http://www.rasch-gz.com>

The student version comes with a small data matrix (30 items x 40 subjects) plus short video. Following the user guide, the user would obtain all the result files at the click of a mouse. This offers good illustration regarding how helpful Rasch model is to the users' field of study or classroom teaching. For equating with tests on large scale yet with high stakes, the users may apply for the professional version of (Rasch-GZ) online.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

List of Abbreviations

COVID-19,	<i>Corona Virus</i> Disease 2019
CTT,	classical testing theory
DI,	discrimination index
ETS,	Educational Testing Service
MC,	multiple choice
MET,	Matriculation English Test
SD,	standardized deviation

References

1. **Zhang, Q.** (2000). BILOG and PARSCALE: Different but Alike. In Jose Lai & Pauline Po-yee Tam. (Eds.). *Crosslinks in English Language Teaching*. Vol. 1, 2000. English Language Teaching Unit. The Chinese University of Hong Kong, Hong Kong, SAR, China
2. **Zhang, Q.** (2004a). [Item Analysis and Test Equating: Research and Application]. [M]. Higher Education Press, China
3. **Zhang, Q.** (2012). Towards International Practice of Language Testing in China. Keynote speech given at the PROMS2012, Jiaxing University, China, August 6-9, 2012.
4. **Zhang, Q & Yang, H.** (Eds.). (2012). Pacific-Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding. Springer. <https://doi.org/10.1007/978-3-642-37592-7>
5. **Zhang, Q & Zhang, T.T.** (2014). Rasch Model: Status quo and prospect in China. In Zhang & Yang. (Eds.). Pacific Rim Objective Measurement

Symposium (PROMS) 2014 Conference Proceedings Rasch and the Future. (pp.17-24). <https://doi.org/10.1007/978-3-662-47490-7>

6. **Zhang, Q.** (Ed.). (2016). Pacific-Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceeding. Springerdoi: 10.1007/978-981-10-8138-5;
7. **Zhang, Q.** (Ed.). (2015). Pacific-Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceeding. Springerdoi: 10.1007/978-981-10-1687-5;
8. **Zhang, Q & Yang, H.** (Eds.). (2014). Pacific-Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceeding. Springer. <https://doi.org/10.1007/978-3-662-47490-7>
9. **Mok, M. M. C., & Zhang, Q.** (Eds.). (2014). Constructing variables. Book of Abstracts Vol. II. Journal of Applied Measurement. ISBN 978-1-934116-10-4. JAM Press. P.O. Box 1283 Maple Grove, MN55311. USA.
10. **Mok, M. M. C., & Zhang, Q.** (Eds.). (2015). Constructing variables. Book of Abstracts Vol. I. Journal of Applied Measurement. ISBN 978-1-934116-11-1. JAM Press. P.O. Box 1283 Maple Grove, MN55311. USA.
11. **Mok, M.M. C., & Zhang, Q.** (2018). [Introduction to Rasch Measurement]. ISBN 978-1-934116-13-5. JAM Press. P.O. Box 1283 Maple Grove, MN55311. USA.
12. **Zhang, Q.** (2019). Rasch Model: Research and Practice in China. In Myint Swe Khine. (Ed.). *International Trends in Educational Assessment*. Brill | Sense. Retrieved from <http://catalog.loc.gov>.
13. **Zhang, Q& Wei, J.G.**(2023). RASCH-GZ (Version 2.0) [Computer software]. <http://www.rasch-gz.com>
14. **Zhang, Q** (2022a). From GITEST to RASCH-GZ: Inheritance and Development of Rasch-based Research in China. <http://dx.doi.org/10.20431/2349-0381.09S1001>
15. **Zhang, Q.** (2022b).Rasch Model and Test Equating in China: The most updated development in China. In Mariam Haji Monek (Chairs), International Education e-Colloquium, the 22nd-24th Feb. 2022. KL, Malaysia.
16. **Zhang, Q.** (2021a). Rasch-GZ, the first Chinese version of Rasch-based item analysis and test equating. In Department of Psychology, Nanjing Normal University (Chairs), Online Symposium. PROMS2021, Nanjing, China. Retrieved from <http://www.proms.promsociety.org>
17. **Zhang, Q.** (2021b).Rasch Model and Test Equating in China: Rasch Model in Chinese version has come. In Xinling Zhang (Chairs), The 6th International Conference on Language Testing and Assessment. Shanghai, China
18. **Thorndike, R., L** (1976). Educational Measurement. 2nd Edition, USA.
19. **Brian K. Lynch.** (2003). Language Assessment and Programme Evaluation. Edinburgh University Press. In Alan Davies & Keith Michel. (Eds.). Edinburgh Textbooks in Applied Linguistics.
20. **P.K. Sivakumaran.** (2007). Further Methods of Correlations. In S.C. Gupta. [Ed]. Fundamentals of Applied Statistics. Sultan Chand and Sons.

21. **Gui, S.C.**, Li, W & Zhang, Q. (1993). [The Application of IRT to MET Equating]. In NEEA (Ed.). (pp. 391-393). The 4th Annual Forum on Educational Testing in Beijing, China Peace Press.
22. **Gui, S.C.** (1986). (pp.84-85) [Standardized Test: Theory, principle and method]. 1st ed., Guangdong, China
23. **Liu, Yuming.** (2020). *Test Equating, Scaling, and Linking: Methods and Practices* (Kolen, M.J., & Brennan, R. L, Trans.; (3rd Ed). (Original work published 2004)
24. **Hambleton**, Swaminathan & Rogers. (1991). *Fundamentals of Item Response Theory*. Newbury Park. California: Sage Publications, Inc.
25. **Hambleton**, Swaminathan & Rogers. (1985). *Item Response Theory: Principle and Application*. Academic Publisher.
26. **Kolen, M.J.,**& Brennan, R. L. (2004). *Test equating, Scaling, and Linking: Methods and Practices*. (2nd Ed). Springer Vertag.
27. **Kolen, M.J.,**& Brennan, R. L. (1995). *Test equating: methods and practices*. Springer Vertag. New York, Inc.
28. **Wright**, B. D., Nead, R.J. & Bell, S.R. (1980). BICAL: Calibrating items with the Rasch model. Research Memorandum 23C. Department of Education, Chicago University.
29. **Wright**, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.
30. **Wright**, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions*, 6(1), 196–200.
31. **Trevor G. Bond**& Zi Yan. (2018). Exporting to China: The future of a Genuine Collaboration with the West. In Zhang Q. (Ed.). Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings Rasch and the Future. (pp.39-48). <https://doi.org/10.1007/978-981-10-8138-5>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

