ISSN 2349-0373 (Print) | ISSN 2349-0381 (Online)

Volume 9, Special Issue 1, May 2022

INTERNATIONAL JOURNAL OF HUMANITIES, SOCIAL SCIENCES AND EDUCATION

Special Issue: Language Testing in China and beyond

Guest Editor : **Zhang Quan**

www.arcjournals.org

ISSN 2349-0373 (Print) | ISSN 2349-0381 (Online)

Volume 9, Special Issue 1, May 2022

INTERNATIONAL JOURNAL OF HUMANITIES, SOCIAL SCIENCES AND EDUCATION

Special Issue: Language Testing in China and beyond

Guest Editor: Zhang Quan

www.arcjournals.org

CONTENTS

PREFACE

ARC Publications is to be congratulated for making this special issue available to language testing professionals, practitioners, teachers, and postgraduate students in China.

It is well known that modern language testing, an important branch of applied linguistics, has been improving with the development of educational measurement in the early 20th century. Over the past nearly a century, the development of language testing has gone through three stages: the pre-scientific stage, the stage combining psychometrics and structuralism, and the one based on psycholinguistics and sociolinguistics (B. Spolsky). Today, in the language testing domain, it is also generally believed that the 1980s witnessed two important shifts: one from classical testing theory (CTT) to item response theory (IRT), indicating that the testing theory has come of age; the other from the use of computer for statistical analyses to the use of computer programming technology for testing administration, showing that testing method has undergone significant changes, and in recent years, from computer-based testing to Internet-based testing, indicating that not only testing method has been greatly updated but the media used in testing have been colorfully varied, also the testing distance has been enormously extended and stretched out. It can be said that research and development in the field of language testing are updating with each passing day.

China is a big country with a long history of language testing. The annual number of candidates taking foreign language tests of various kinds is also the largest in the world. However, there have been existing controversial issues for its measurement attributes, validity, washback effects, etc. Or rather, there exists vast space between testing theory and actual practice. Therefore, what our journal needs to focus on is no longer the issues regarding how to estimate a certain parameter more accurately, or whether the data fit the model or vice versa, but whether our testing methods and means are being practiced more correctly.

The present issue is a representative selection of eight up-to-date articles, covering a wide range of theoretic, vocational and corpus subjects ad hoc devoted to language testing and falling into five categories: feature articles, testing theory, testing and teaching, testing software, culture and society, and conference reports.

Among them, Chapter One is a feature article, introducing from a diachronic point of view, GITEST, the only Rasch-based testing software in China originally programmed in BASIC language, running under DOS that can handle item analysis and test equating of the small-scale data matrix, after decades of use and moderation, has been fully upgraded to RASCH-GZ, capable of processing infinite data matrix, of running on the network with online technical support, and of being well geared with today's international practice. The article shows, from GITEST to RASCH-GZ, the inheritance and development of language testing in China.

Chapter Two is a series of reports from Chinese conference organizers. Starting from PROMS 2021, Nanjing, China, with beautiful photos and detailed presentation, the author summarized PROMS conferences held in China over the past ten years (2012-2021), informative and beneficial to scholars and testing counterparts home and abroad.

Chapter Three addressed that course examination, as an important link in the teaching of higher vocational education (HVE), plays a key role in cultivating talents in China. The author specified four factors: difficulty level, proportion of non-standard answers, proportion of classroom grades and proportion of professional practice questions in the course examinations and analyzed using orthogonal test method via SPSSAU. The results can be used as an important

reference for increasing the rationality of test content, improving evaluation, optimizing test management, and enriching testing methods.

Chapter Four and Chapter Five were dealing with testing software and practice. The former is a preliminary research and application of WINSTEPS and GITEST. Basic concepts and specific methods of test equating were fully discussed. The author concluded that WINSTEPS and GITEST are different yet alike and both are equally effective for test equating. Meanwhile, the authors also verified the hypothesis originally proposed by Wright and Stone (1979): the linking items are the "HARD" items in the EASY test but the "EASY" items in the HARD test. The author of the latter article introduced, from the perspective of computer programming technology and user guide plus nice illustrations, RASCH-GZ, the First Chinese Version of Rasch-Based Item Analysis and Test Equating System. Rasch-GZ was successfully developed during the COVID-19 epidemic period. This updated RASCH-GZ in recent days provides Chinese scholars with a powerful help in popularizing the research and application of the Rasch model, showing the inheritance and development in Rasch-based research for language testing in China.

Corpus is also one of the essential tools used in language testing. Chapter Six addressed a three-level corpus study. By examining the text complexity of reading comprehension passages in China's National Matriculation English Test（NMET）of Year 2020 and 2021, on the purpose of providing validation evidence for this new NMET reform, the author showed that MET in China required a much larger vocabulary size than the number indicated in the guidelines, and more often, of thematic context and genre and that the passages of the two-year NMET employed unproportioned use of human and society and exposition. Good references for Chinese counterparts.

In addition to CTT, Rasch model and corpus study, the author of Chapter Seven presented "The Impact of Social Presence on Learning Satisfaction of Chinese learners online". Today in China, this is an important yet popular topic in online learning on campus. The author analyzed the effects of social presence on cognitive presence, emotional presence, and learning satisfaction, and attempts to empirically analyze whether learning readiness has a moderating effect on these relationships. Both EFA and CFA were used. The conclusion was satisfactory. Chapter Eight gave a different perspective, concentrating on an investigation about the variables of feedback seeking behavior (FBS) and job performance in the cross-cultural work settings. The empirical study was conducted in the multinational companies located in China. This research is innovative in theory and methodology which enriches managerial literature by exploring more inclusive solution that benefit work outcomes in cultural diversity teams.

The availability of such a broad range of materials will greatly help readers realize the status quo and development of language testing in China and beyond. Meanwhile, it will also contribute to building up language testing into an independent discipline in Chinese universities and colleges.

We very much appreciate the ARC Publications for providing us with such a wider platform.

Editor

Zhang Quan

Guangzhou, China

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).

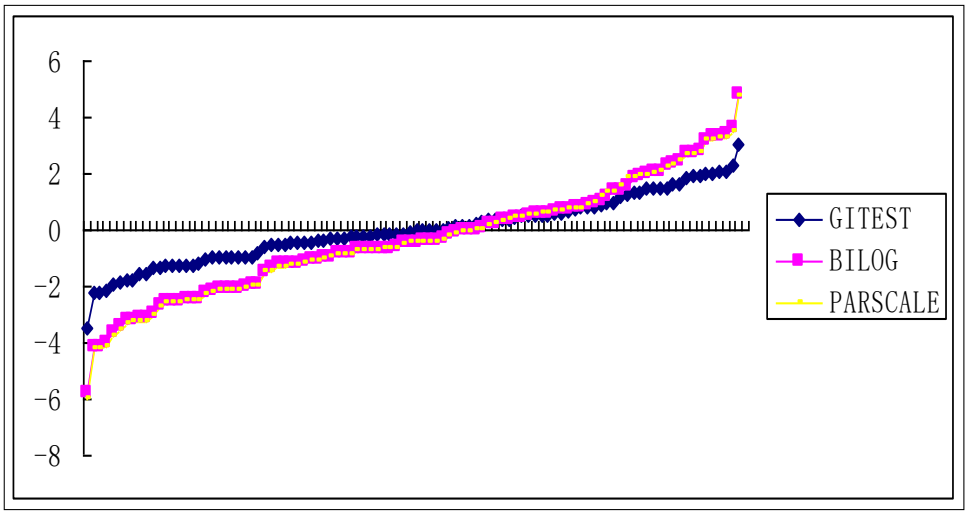


Figure1. *Difficulty curves of GITST, BILOG and PARSCALE based on the same MET data (1990-1999)*

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%.

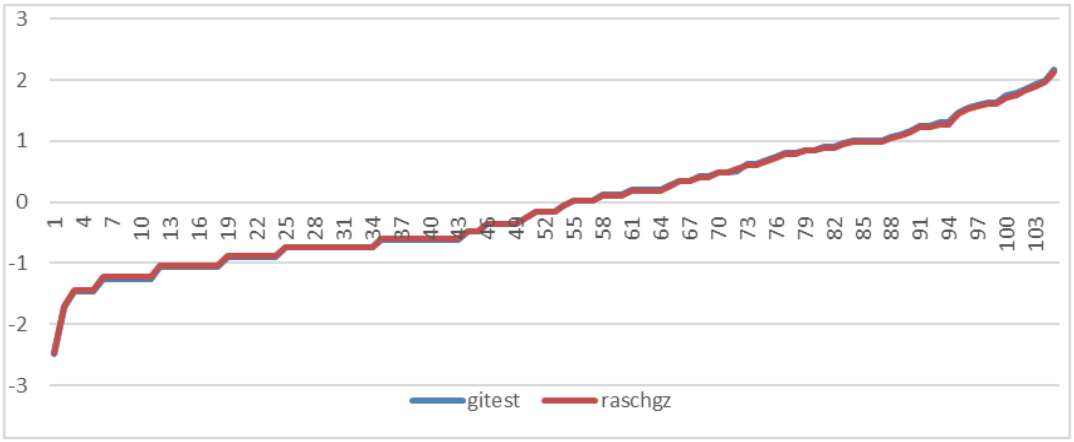


Figure2. *The difficulty curve of the test items generated by GITEST and Rasch-GZ on the same data processing*

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.*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 8-14
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1002
www.arcjournals.org

An Overview of the PROMS in China ----- Retrospect and Prospect from Chinese Organizers

Hu Xiaoxi

Southern Medical University, Guangzhou, China

Liu Ting

Guangdong Polytechnic of Science and Technology, Guangzhou, China

Corresponding Author:** ***Hu Xiaoxi, *professor of Southern Medical University, Guangzhou, China, is both the successful organizer yet active sponsor of PROMS 2014 and 2016, China. sheila126@ 126.com. Liu Ting, PhD of City University of Macau, SAR, China, is actively involved in social media's impact on tourism and education.*

1. INTRODUCTION

The Pacific-Rim Objective Measurement Symposium (PROMS[1]) 2021 just held in Nanjing was successfully concluded. This is the first large-scale online international seminar successfully held in China at a difficult time in the fight against the COVID-19 epidemic under the global environment. PROMS 2021 was jointly organized by the School of Psychology, Nanjing Normal University, China and the Jiangsu Provincial Psychological Society, and was realized on the online conference platform ZOOM. This year more than 100 experts, researchers and scholars participated. They are mostly in the field of psychology and education from 11 countries and regions including the United States, Australia, Brazil, Sweden, Japan, Singapore, Malaysia, Vietnam, Hong Kong, Macau and mainland China.

A one-day pre-conference workshop was run. The workshop was divided into two parallel series. Series (1) "Rasch Model Workshop", conducted by Dr. Yan Zi, Associate Professor of the Department of Curriculum and Teaching of the Educational University of Hong Kong and Vice Chairman of the PROMS. The main takeaway is that the Rasch model focuses on items and people rather than on test scores, using the joint measurement principle to synthesize the two; quantitative analysis of qualitative questions is now experienced in a way rarely practiced in the social sciences. The workshop explains the advantages of Rasch-based analysis over classical test and questionnaire scoring methods. How Rasch model-based data analysis can be applied to dichotomous rating data (basic Rasch model) and Likert-style questionnaire data (Rasch model rating scale model).

The workshop series (2) is led by two professionals, the "R-language Workshop" by Dr. Cynthia Tong from the Department of Psychology, University of Virginia, USA, and "Educational Data: Storytelling and Ways to Find Teaching Implications" workshop by Dr. Kit Tai HAU, a professor of the Department of Educational Psychology, The Chinese University of Hong Kong, China. Among them, "R Language Workshop" mainly introduces what R is and how to apply it in Rasch model measurement. This workshop is useful for scholars interested in T-tests, ANOVA, multiple regression. Confirmatory factor analysis (CFA), measurement invariance, and structural equation modeling (SEM). The workshop by Dr. Hou deals with two aspects: (1) How to select the questionnaires used in education monitoring; (2) How to use the results of the questionnaires to analyze and write attractive, Influential "Stories". Dr. Hou used PISA-type tools and their materials to illustrate and serve as examples. Workshops focused on training in lively and interesting "stories" rather than on complex and sophisticated statistical techniques.

The theme of PROMS 2021 is: Application of Rasch Model in Psychological and Educational Research. The organizer invited four experts to make keynote speeches.

[1] http://proms.promsociety.org

They are Professor Trevor Bond of James Cook University, Australia, the former president of PROMS, who gave a keynote speech titled: From Estimation to Consideration: The Role of Rasch Model Measurements in Promoting Understanding and Improving Validity; Professor Dr. Kit Tai HAU, whose keynote focused on "Large-scale International Educational Assessment: Uses, Limitations, and Counter-Intuitive Findings; Professor Ricardo Primi from the University of São Francisco, Brazil, whose speech is about "Reactive Styles as a Function of Individual Differences: Addressing the Multiplicity of Individual DIFs, and Professor Steven Stemler from Wesleyan University, USA. Prof. Steven delivered the keynote titled: "Better Measurement and Fewer Parameters! The True Value of Rasch over IRT!

In addition, the conference also invited Professor Zhang Quan, the PROMS mainland China representative from Jiaxing University, Zhejiang Province, China to run a symposium called "Rasch Model and Test Equating: RASCH-GZ, the most updated research in China". This shows to the academic community the inheritance and development of Rasch model research in China as well.

2. INTRODUCTION TO PROMS

PROMS was initiated and chaired by Professor Trevor Bond. The "Pacific Rim Objective Measurement Symposium" hosted by the society is referred to as PROMS in the world. It is the annual official academic conference of PROMS. The purpose thereof has three main points:

- Encourage the pursuit of Rasch model measurement and its applications in business, consulting, economics, education, healthcare, languages, measurement, psychology, quality assurance, statistics and strategic planning in the Pacific Rim;
- Advocate for measurement practices that contribute to the well-being of individuals, communities and societies, and the public interest in the Pacific Rim;
- Increase the visibility and status of the Rasch model measurement in the Pacific Rim.

These Goals will be Achieved by:

- Building a network of researchers and scholars in the Pacific Rim to foster collaboration and collaboration;
- Holding an annual conference-----PROMS;
- Providing seminars and training;
- Participating in other activities that may contribute to the development of Rasch model measurement research, policy and practice.

The PROMS Conference is a forum for sharing new knowledge and new applications of the Rasch model with the international academic community, which is embodied in the application of the Rasch model in various academic fields, aiming to contribute to the promotion of objective measurement research and development in the Pacific Rim region.

Since the first PROMS was held in Kuala Lumpur, Malaysia in 2005, it has been held for 15 consecutive academic years: Malaysia (2005, 2010, 2017), Singapore (2011), Japan (2008, 2015), Indonesia(2019), China Hong Kong (2006), China Taiwan (2007, 2013), China Jiaxing (2012), China Guangzhou (2014), China Xi'an (2016), China Shanghai (2018) and China Nanjing (2021).

The PROMS conference has the following characteristics: It is held once a year in the Pacific Rim region. The participants are all MA or Ph.D. supervisors, scholars, and MA or Ph.D. students who are engaged in Rasch model research and objective measurement or testing. The fields of their research can be different, but they all use the Rasch model to do the research of their own. The software systems used include WINSTEPS, FACET, GITEST, etc. to process their own data and exchange research results. Apart from inviting influential professionals and scholars in the field to deliver keynote speeches and arrange parallel sessions, the organizer(s) of each conference also arranges, according to international practice, a 1-2-day pre-conference workshop on the promotion for the knowledge of as well as the use of Rasch-based software to popularize the research and application of the Rasch model for beginners, graduate students, and young teachers. It has been well believed that the PROMS conferences provide an excellent academic platform for graduate students, supervisors and scholars in the Pacific Rim and beyond.

PROMS 2013 Kaohsiung, China

3. THE ORIGIN OF THE RASCH MODEL[2]

The Danish mathematician Georg Rasch (1901-1981) published "Probabilistic Model for Intelligence and Attainment Tests" in 1960, which resolved the debate "Is psychometric possible?" It is generally believed that physicists and social scientists had hot debating in the 1930s. Georg Rasch's publication shows how the rigorous standards of physical scientists can be applied to the social sciences through the model Rasch proposed. This became known as the "Rasch model". Since then, the Rasch model has been widely used in various research fields such as educational measurement, sociology, management, psychology, language testing, medicine, and health sciences, and have been extended from their initial application to data of dichotomous type to polychotomous type. Today, the model has been respected as the "Rasch model" in the academic circles. Experts and scholars are also convinced that this model has made a great contribution to the promotion of the prospect of objective measurement and scientific exploration. And PROMS conferences are the most authoritative, professional and constantly updated international conferences on Rasch model and their theories and applications.

At a technical level, among the many measurement methods, the Rasch model can be regarded as the most versatile, powerful and feasible. Whether it is data matrix of binary scoring or multi-level scoring, it can be processed via Rasch-based software. With just one click of the mouse, the data results to be processed can be stored in a designated computer file for easy access at any time in the future. However, most of these widely used, powerful and feasible methods are limited to the English-speaking world. Most researchers who are native Chinese speakers and non-English majors may not fully understand and feel the concepts, applications and efficiency inherent in these methods. For quite a long time, it has been the situation of liberal arts in academic circles in China.

At the theoretical level, as far as language testing is concerned, the classical testing theory is still the main measurement model for schools and examination departments in China. This is reflected in the fact that the difficulty of test items depends on the level of the subject group, and the level of subjects restricts the difficulty of test items in turn, while the advantage of Rasch model is that the difficulty of

[2]http://www.rasch.org/rmt/rmt193h.htm

the test items and the test taker's ability level can be estimated independently of the sample (Andrich, 2004; Bachman, 1996; Engelhart, 2013; Fan, 1998; Gui, 1986, 1990, 1993; Hambleton & Swaminathan, 1985; Kolen, 1995, 2004; Linacre, 2002, 2016; Smith, 1990; Trevor, 2015;Wright, 1979, 1992). In this sense, the Rasch model is an effective method for language testing research. This is also the author's original intention to promote the research and development of the Rasch model in China.

4. FEATURES OF PROMS WORKSHOPS

The PROMS workshops have three characteristics: first, the content basically remains unchanged, and it is all about popularizing the application of the Rasch model. This includes the introduction to and the specific use of WINSTEPS, FACET, LEXILES and TAM, R system and EQSIRT, etc.; next, the speakers are all influential experts and professors in the Rasch research field; thirdly, the lecture format is very precisely targeted. Considering that many of the learners are non-English majored teachers and graduate students, both Chinese and English instructions are adopted. Very often the question and answer as well as relevant discussions were mainly in Chinese. The purpose is obvious: to better promote the understanding and application of the Rasch model among Chinese scholars.

In fact, PROMS entered China mainland as early as in 2012, and was successfully held for the first time in Jiaxing University, Zhejiang province, China. Since then, the organizers of the conference have invited Professor Trevor Bond to give workshops for six consecutive times (Jiaxing, 2012; Kaohsiung, 2013; Guangzhou, 2014; Xi'an, 2016; Shanghai, 2018; Nanjing, 2021), the content of which is: "How to use WINSTEPS to conduct Rasch Model Measurement". This is actually a very popular class on how to apply the Rasch model for research and data processing and analysis in China as well as in the Pacific rim region. It should be pointed out that in China there are still not many people who can use the above-mentioned software skillfully. Therefore, such a popular workshop is necessary to continuously promote the study and application of the Rasch model. For Chinese scholars and researchers in their respective research fields, Rasch model has played a positive guiding role in the practical application.

5. PROMS THEMES AND KEYNOTE SPEAKERS

The PROMS 2021 held in Nanjing is the sixth PROMS conference in China. The PROMS topics are all about the "Rasch Model: Theory, Method and Practice", and the research fields involved range from medicine, education, sports, network technology, language teaching, language testing, validity research and other fields. From 2012 to 2021, approximately 50 keynote speakers have been invited from universities in Malaysia, the United States, Australia, Denmark, Japan, Brazil, Sweden, Singapore, Vietnam, and China including Hong Kong, Macau, and Taiwan.

The keynote speakers are all centered on the Rasch model. It is worth mentioning that the PROMS conferences held in China have also tested the qualification of Chinese scholars in the research and application of the Rasch model. Among them, in 2012, Professor Zhang Quan of Jiaxing University gave a keynote speech to report to the international academic community, for the first time, how the late famous Chinese linguist Professor Gui Shichun (1933-2017) successfully completed the ten-year MET equating project via Rasch model sponsored by the National Education Examination Authority (NEEA) under the Ministry of Education of China. In 2014, Prof. Jin Yan from Shanghai Jiaotong University and Dr. EricWu from UCLA jointly made the speech titled: "An Argument Approach to Test Fairness: The case of multiple-form equating in the College English Tests. In 2016, Professor Huang Xiaoting of Peking University gave a keynote speech titled: "Investigating the Predictive validity and Social Consequences of “Gaokao”. These speeches made the participants refreshing and impressive. Generally speaking, the convening of the PROMS Conference in China has enabled the academic community to have a better understanding of the time and level of Chinese scholars’ application of the Rasch model.

6. CONCLUSION

To conclude, PROMS conferences held in China offer three highlights: First, with the number of PROMS participants increasing annually, the academic impact becomes obvious in the Pacific rim region. At the same time, it shows to the academic community that Chinese scholars have inherited and developed the Rasch model research. And it also symbolizes that the research on the Rasch model by Chinese scholars has entered into the world.

Secondly, over the past years, the PROMS has been concerned, supported and sponsored by many work units home and abroad: the world-renowned Springer Publishing House, the US Meta Metrics data company, the ETS, the British Embassy Cultural and Educational Association, SAGE, Higher Education Press, Social Media, Foreign Language Teaching and Research Press, Shanghai Foreign Language Education Press, Jiaxing University, Shanghai Jiaotong University, Fudan University, City Training Institute CTI Guangzhou, Jiangsu Psychological Association (Development and Educational Psychology Special Committee). “Correction Network” and related institution Hangzhou.

Thirdly, the high quality of the papers presented at the conference reflect the new trends and status quo of the specific research and application of Rasch model in the Pacific Rim region. For four years Springer Publishing officially published the PROMS Proceedings (Zhang & Yang, 2012; Zhang & Yang, 2014; Zhang, 2015, 2016). Ever since the publication, the paper-based copies of the conference proceedings have been collected in major libraries and universities around the world, and the e-books are included in the Springer Link Behavioral Science and Psychology e-book collection, which is accessible to readers around the world. The annual number of paper downloads and citations increase year by year, which has a great impact in the field of Rasch research.

With the advent of the big data era and the further improvement of computer and network technology, the research and application of the Rasch model will be popularized in more research fields in China. In the forthcoming PROMS conferences, Chinese scholars are expected to further enhance academic research and communication skills and to better manipulate the direction of research and development of their own disciplines. The annual PROMS will continue to add vitality to language testing and objective measurement research in China as well as in the Pacific Rim region.

REFERENCES

Andrich,D.(2004).Controversy and the Rasch model: A characteristic of incompatible paradigms? Medical care, 2004, 42 (1):1-7

Bachman, L. F. & Palmer, A.S. (1996) Language testing in practice. Oxford University Press.

Engelhard, G. Jr. (2013). Invariant measurement using Rasch models in the social behavioral, and health sciences. Routledge Taylor Francis Group. ISBN: 978-0-415-87122-8 (hbk). ISBN: 978-0-41 5-8712S-9 (pbk). ISBN: 978-0-203-07363-6(ebk)

Fan,X.(1998).Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement.

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

Gui, S.C. (1990). Notes on Itembanking (2).Modern Foreign Languages. Issue 4, pp.66-72.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principle and application. Kluwer Nijhoff Publishing.

Kolen, M.J. & Brenman, R L. (2004).Test equating, scaling and linking methods and practice. Springer Verlag New York Inc.

Kolen, M.J. & Brenman, R.L. (1995).Test equating: method and practice. Springer Verlag New York Inc.

Linacre, J.M. (2016).A users guide to WINSTEPS Rasch model computer programs: Program manual 3.92.0[EB/OL] Retrieved from https://www.winsteps.com/winman/copyright htm, 2016.

Linacre, J.M (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions 878, 2002.

Smith, R.M. (1990). Theory and practice of fit. Rasch Measurement Transaction, 1990, 3(4):78.

Trevor, G.B., & Christine, M Fox (2015).Applying the Rasch Model: Fundamental Measurement in the Human Science. (3rd Ed.).Routledge Taylor Francis Group.

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

Wright, B.D. (1992).IRT in the1990s: Which models work best? Rasch Measurement Transactions, 1992, 6(1): 196-200.

Zhang, Q. (ed.). (2016). Pacific rim objective measurement symposium PROMS 2016 Conference Proceedings. ISBN: 978-981-10-8137-8; ISBN: 978-981-10-8138-5(eBook). doi: 10.1007/978-981-10-8138-5

Zhang, Q. (ed.). (2015). Pacific rim objective measurements Symposium PROMS 2015 Conference Proceedings. ISBN: 978-981-10-1686-8ISBN:978-981-10-1687-5(eBook). doi: 10.1007/978-981-10-1687-5

Zhang, Q &Yang, H. (eds.). (2014). Pacific rim objective measurement symposium PROMS 2014 Conference Proceedings. ISBN: 978-3-662-47489-1; ISBN: 978-3-662-47490-7(eBook). doi: 10.1007/978-3-662-47490-7,2

Zhang, Q & Yang, H. (eds.). (2012). Pacific rim objective measurement symposium PROMS 2012 Conference. Proceedings. ISBN: 978-3-642-37591-0; ISBN: 978-3-642-37592-7 (eBook). doi: 10.1007/978-3-642-37592-7,2

Citation: *Hu Xiaoxi and Liu Ting. "An Overview of the PROMS in China ----- Retrospect and Prospect from Chinese Organizers" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 8-14. DOI: https://doi.org/10.20431/2349-0381.09S1002.*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 15-24
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1003
www.arcjournals.org

Towards Better Evaluation on Course Examination of Tourism English Major for Higher Vocational Education in China

Liu Jiajie

Hebei Institute of International Business and Economics China.

Corresponding Author:** ***Liu Jiajie, *Associate Professor, PhD. in International Tourism Management, City University of Macao, is director of the Tourism English Teaching and Research Section, Hebei Institute of International Business and Economics in China. Dr. Liu's research interests involve tourism management, tourism English testing and teaching. 1025362761@ qq.com*

Abstract: *Course examination, as an important link in the teaching of higher vocational education (HVE) plays a key role in cultivating talents in China. This study used the humanistic pedagogy as a humanized method to evaluate the degree of satisfaction for the course examination. Four factors: difficulty level, proportion of non-standard answers, proportion of classroom grades and proportion of professional practice questions in the course examinations were specified and analyzed using orthogonal test method via SPSSAU. The experiment data were collected in the course examinations for tourism English major of Hebei Institute of International Business and Economics in China. The optimal combination of various factors in the course examination evaluation was obtained. The results can be used as an important reference for increasing the rationality of test content, improving evaluation, optimizing test management, and enriching testing methods.*

Keywords: *Higher vocational education (HVE), Tourism English major, Course examination, Orthogonal test method (OTM)*

1. INTRODUCTION

Higher Vocational Education (HVE) in China is an integral part of higher education. It is different from the HVE system practised in Australia, the dual system in Germany, and the community education system in the United States. To quote some Chinese experts, in China, vocational students are full-time and school-age youth, with the goal of serving the society (Ma, 2010). By China's HVE is referred to higher education activities implemented to enable the students to obtain the field expertise the relevant knowledge of science and culture, professional ethics and technical skills required for a certain occupation or career development.(Institute of Vocational and Technical Education Center of the Ministry of Education, China 2016). The purpose of HVE in China is to cultivate talents for specific applications and workers with certain professional knowledge and skills (Liu, 2019). Compared with higher education in China, HVE pays more attention to the cultivation of practical working ability and skills. Therefore, the training goals of Chinese higher vocational colleges are both career and employment-oriented, which in turn determines that all aspects of education and teaching within HVE must closely focus on the core content of career development (Guo, 2015).

Examination is an important means of educational evaluation, one of the important links in the implementation of HVE, and a concentrated performance in evaluating the quality of student training (Wu, 2011). In particular, the professional course examination is the core basis for evaluating students' performance. It can not only measure students' learning effects but also reflect students' mastery of knowledge and content stipulated in the curriculum. It can evaluate teachers' teaching quality as well. What characterizes the course examination is of non-placement, diagnosis, criterion-referenced, normal distribution and relativity in nature (Xie, 2021). Here in the author's opinion, through curriculum, students have learnt understood the curriculum knowledge, transformed the learning into a knowledge system according to their own cognition, formed self-experience and concepts, and built up the schema of their own. As the evaluation of student performance is based on the core content learnt, the test scores should reflect students' ability in terms of learning and of

comprehensively applying the course knowledge. Apart from this, students should also be encouraged to actively participate in the whole process of course teaching and to think, to check and fill in the gaps in time and thus to enhance the mastery of course knowledge. Meanwhile, students' ability regarding comprehensive thinking and flexible knowledge application would also get greatly enhanced. In this sense, to establish an objective, standard, scientific yet operational evaluation method for course examination is beneficial to improve the quality of course examination. In this way, both students' mastery of knowledge and improvement of teaching quality can be strengthened. It is based on these ideas that the author addresses the better evaluation on Course Examination of Tourism English Major (CETEM) for HVE in China.

2. LITERATURE REVIEW

2.1. Examination

It is known to all, standardized examination was first introduced into China by the late Chinese famous linguist Professor Gui Shichun (1930-2017) who is also the first professional conducting successfully the ten-year (1990-1999) Matriculation English Test (MET) equating project sponsored by Ministry of Education of China (Gui,1986, 2007, 2017; Li, 2000; Zhang, 2014, 2016). For the definition of the test, the academic circles have different interpretations. According to Sheng (1997) in "Ancient China's Examination System", "examination" was defined as "accurate measurement of talent's knowledge and ability." In the "Chinese Educational Dictionary", examination is etymologically explained as "one of educational measurement tools, allowing the test takers to answer test questions, according to certain assessment purposes, in line with a compulsory manner within a given time period, and grades and scores were given based on the test takers' final performance" (Gu, 1998). The Western Education Dictionary defines an exam as a means of assessing in written or spoken mode a test taker's quality of thinking, knowledge, and ability (Rowntree, 1988). With the development of the Internet, computer technology and big data, the format and test item types have undergone great changes. Computer-based exams and internet-based exams have increasingly higher requirements for the construction of item banking. This paper adopts a narrow definition of examination, i.e. an exam is referred to as a kind of social activity wherein certain or multiple aspects regarding examinees' quality or proficiency in a certain organization were tested, screened or evaluated with selected resources, according to the examination purposes (Liao, 2003). The course examination in question to be addressed in the present article falls into the achievement test in terms of the examination purpose.

2.2. Course Examination

Curriculum was derived from the Latin word "currere", originally meaning 'runway'. Since modern times, people have had great differences in the definition of curriculum. So far no consensus ad idem has reached. In the author's opinion, the concept of curriculum can be roughly divided into four categories. The first category defines courses as teaching subjects. The second one interprets the curriculum as a kind of learning experience. The third one considers the curriculum as cultural reproduction. The fourth one takes the curriculum as a process of social transformation. According to a Chinese scholar (Xu, 2015), many Chinese teachers and students agreed upon with the first concept. However, such a concept neglects the feelings of students, thus sort of confined. The second one was put forward by American educator named John Dewey who believed that the curriculum is simply the experience or rather the experience that students gain spontaneously or under the guidance of teachers. Students' learning depends on what they do, not on what teachers instruct. The highlight here is placing students' direct experience at the center of the curriculum. Dewey took into account the importance of students' active acquisition of experience, but in some way failed to value the learning of professional knowledge and systematic knowledge; therefore, the definition of this kind remains limited. The third and fourth concepts only explain the curriculum from a specific point of view and cannot totally cover all the connotations of the curriculum. The curriculum addressed in this paper includes both teaching subjects and students' feelings. It is the sum of all the items put together including the disciplines and majors prescribed to achieve the teaching objectives of higher vocational

colleges, training purposes, training content, talent training models and processes, and students' feelings.

Course examination in higher vocational colleges (CEHVC)are a kind of social activities administered by the teaching managers of various majors to measure and evaluate the student ability by using relevant resources according to both the professional characteristics and the needs for talented personnels. Such an exam is different from the course examination for general higher education (CEGHE) in China. Compared with the uniformity of the latter, CEHVC is more flexible and open. The functions of both examinations are the same in terms of evaluation, motivation, guidance, feedback and adjustment. Here three points need stressing: at the first place, exams play the role of a baton, directing students in their studies, especially for those students who are strongly motivated to get higher scores (Zhou, 2015). Secondly, it can evaluate students' overall and phased learning outcomes. Third, based on the achievements made by students, the teaching managers can obtain the feedback concerning weak links, quality of classroom instructions, and shortcomings in the teaching process. According to the feedback, on the one hand, teachers can make corresponding improvements. On the other hand, students can adjust their way of learning and thus fill in gaps. In addition, the results of the exam will motivate students to keep studying harder and be better prepared for next exam(s).

2.3. Course Examination of Tourism English Major (CETEM)

Testing purpose falls into four types: proficiency test, achievement test, aptitude test and diagnostic test (Gui, 1985). The achievement test is used to measure how well students completed a syllabus or textbook within a period of semester. Achievement checks up learning progress. Such a test must echo the teaching over the past, and the test items designed or the test contents used must be based on what was taught in the classroom or the textbook used. The score on the achievement test can indicate the degree to which the student has mastered the teaching content. If the percentage system is adopted, getting a score of 60 means that students have mastered 60% of the teaching instruction. The score obtained from achievement test is often determined by the following two points:

1) Students’ scores hinges on the amount and the quality of the teaching instruction. The better, the teaching, the higher, the score or vice versa;

2) Intercollegial scores of achievement tests are uncomparable because tests were based on different text books, different schedules, and different ways of testing. Assessing students' proficiency based on achievement test scores is often inaccurate because there is a lack of uniform standards. Judging from the score, the pass rate of the achievement test should be relatively large, and too low a pass rate will bring problems to schools and society. In this sense, the achievement test is a local test confined within a school, and teachers often make their own test items based on the content taught. The annual mid-term, final examinations, placement test and tests for graduation are all achievement tests in nature (Gui, 1985).

CETEM is achievement test. The principles regarding reliability, validity, discrimination, and feasibility are the principles that language testing should follow (Liu, 2015). Graduates majoring in tourism English enjoy more opportunities for job hunting. They can work as tour guides, in tourism management offices and the sections of hotel services as well. Therefore, tourism English majors in higher vocational colleges need to emphasize practicality and professionalism. The content of professional course examinations is closely related to the needs of occupational tasks, highlighting the practicality of and pertinence to English language training. In the present study, the author selected the compulsory course "Hotel English" for tourism English major as the course for testing. The reason is that this curriculum bears obvious features, integrating hotel English vocabulary, common sentence patterns, and professional knowledge. This can fully reflect the teaching effect and students' learning ability.

2.4. Humanistic Pedagogy

Humanistic pedagogy has been practiced since ancient times, from Greek school of sophists comendating that "man is the measure of all things", to Socrates' "know yourself", to Plato's opposition to forced learning, all of which reveal a humanistic flavor. In ancient China, the educator

Confucius also mentioned the humanistic ideas of "tailored teaching according to student aptitude" and "equal teaching to all walks of life". Though the humanistic educational thought has been handed down generation by generation, what really represents this educational theory is the humanistic psychology proposed by the modern scholar Rogers (2015).Based on humanistic pedagogy, Rogers' humanistic education thought is reflected in all aspects of his theory and is the center of his educational thought. Rogers believes that students need to be placed at the center of teaching and that traditional education is useless and ineffective because it made students unable to master the knowledge and have the sense of being a failure (Rogers, 2015). Apart from this, Rogers divides learning into meaningful learning and meaningless learning (Rogers & Freiberg, 2015). Meaningless learning is indoctrination learning that does not involve personal experiences and emotions (Rogers & Freiberg, 2015) and neither does it have much impact on the individual. Meaningful learning, on the other hand, involves personal engagement while stimulating intrinsic interest in the learner and requiring self-assessment by the learner. The reason is simple: the learner knows best whether this kind of learning meets his needs, whether it helps to get what he wants, and whether he understands some aspects that he did not know well. Today, HVE needs a new student-centered perspective and model and regards students as the main participants in educational reform (Wang, Zhong & Xiao, 2020). The idea that humanistic educators attach importance to students is precisely in conformity with the educational goal of higher vocational colleges. Examination, as a link in education, also bears the responsibility of cultivating talented personnel, and its goal is subordinate to the goal of education. Hence, the educational goal of the student-centered thought is the same as the goal of the CEHVC. Therefore, the educational goal of humanistic pedagogy is the same as the goal of the former. In addition, the examination not only has the function of evaluation and diagnosis, but also increases students' interest in and attention to the curriculum. It can help professional courses to better accomplish their training goals. In order to reflect the humanistic pedagogy, the present study used the student's degree of satisfaction with course examination as an outcome measure.

2.5. Satisfaction

Satisfaction is a relatively matured concept that has found a wider application to research fields such as economics and etc. In recent years, the educational circle has also attached great importance to the study of satisfaction. Satisfaction, also termed as customer satisfaction, refers to "the customer's feeling that their requirements have been fulfilled "(Chen, 2010). Such a research can be specified as the students' feelings about the satisfaction with the course examination according to their own needs. The essence inherent in satisfaction theory is a kind of quality view in nature, referring to a kind of value in which the "human" factor replaces the "material" factor (Chen, 2010). Satisfaction theory highlights the value of people, making people the standard to measure things, which can, in turn, fully reflect the humanistic pedagogy. Therefore, this study applies satisfaction based on students' feelings as the result index of the evaluation of the tourism English major in vocational colleges in China.

3. RESEARCH METHODS

3.1. Orthogonal Test Method (OTM)

This research used the orthogonal test method to evaluate the course examination of tourism English in higher vocational college (CETEHVC), in order to better understand the test effect. As we know, of many kinds of experimental design methods to study multi-factor and multi-leve, OTM is believed to be one of the methods to achieve efficient, rapid and economical effects (Xie, 2021). OTM is not only used in social sciences, but also widely used in natural sciences. It is a design method that uses orthogonal tables to arrange multi-factor experiments and analyze experimental results (Liu, 2012). It selects some representative levels and factors from all levels of experimental factors for testing and analysis, and finds out the combination at optimal level so as to achieve the guiding purpose of practical work. The algorithm of orthogonal experimental data is the key to the experiment (Gong & Li, 2004). The biggest feature of OTM is that some tests can be selected to replace all the tests, which is parsimonia yet practical. By testing some tests and analyzing the results, the optimal combination level of influencing factors can be found out so as to understand the situation of comprehensive tests. The evaluation of CETEHVC is composed of a variety of complex factors, often involving fields such

as education, statistics, psychology, economics, and management. This experimental method has significant advantages in solving the evaluation problems of course examinations with complex factors.

3.2. Research Design

To illustrate humanistic pedagogy, the test index of this experiment is the satisfaction rated by students. The student satisfaction rate was calculated by satisfaction. Firstly, a survey on the satisfaction of the course examination was carried out among the students majoring in tourism English. Then, the satisfaction ratio of each test group was calculated according to the weighted average method. Based on previous research and the composition of previous course examination evaluations, the four factors were specified as the key independent variables. They were the difficulty level (DL) , the proportion of non-standard answers (NSA), the proportion of classroom grades (PCG) and the proportion of professional practice questions (PPQ) in the course examinations. It was assumed that there exists no interaction among the four factors. This experiment adopted the orthogonal table of four factors by three levels. SPSSAU system was used for data analysis.

The course examination was divided into two forms: paper-based test and vis-à-vis interview. The paper test was a closed-book test and all the students took the test at the same time. The interview was made up of questions for NSA and PPQ. The questions were determined by lottery, and candidates were randomly selected on the spot. The paper test came from the three-year "Hotel English" course examination papers in 2016, 2017 and 2018.The difficulty parameters of the test papers are 0.69, 0.63, and 0.52 respectively. According to the previous difficulty values, three difficulty levels of the course examination were determined: above average (AA), on average (OA) and below average (BA).These difficulty levels were only reflected in the content of the paper test. The questions for NSA were professional topics for discussion, with 12 optional questions. Each student selected one to answer, and communicated with the interviewer for 1-2 questions. The time given was within 3-5 minutes. The PPQ concerns hotel service project with 12 optional questions. Each student drew a question for a scenario simulated demonstration. The answering time was within 5-10 minutes.

3.3. Research Purpose

The purpose of this experiment was to select the best content of CETEHVC, and to provide reference for the rationality of the examination. Tourism English major is a major with strong language application and practice. In this study, the DL, the NSA, the PCG and the PPQ were selected as the key influencing factors, and their proportion in the course examination was appropriately increased for experiments. This can fully reflect the humanization of teaching evaluation, cultivate students' ability to use curriculum knowledge independently, and give full play to the main role of students in course examinations. The three overarching research questions proposed in the present research are:

(1)What is the influence of each factor on the course examination?

(2)What is the primary and secondary relationship among the influences of various factors on the satisfaction of the course examinations?

(3)What is the best combination of test content for vocational English majors in vocational colleges?

3.4. Participants

Totally, 90 Chinese students (26 males and 64 females) majoring in tourism English of Hebei Institute of International Business and Economics participated in this experiment. They were divided into 9 groups, each group having 10 people. The test curriculum was "Hotel English", the core curriculum for tourism English major. Taking into account the sample homogeneity and representiveness, this study only made case studies for a specific range. One thing worth mentioning is that the English major has been ranked among the top 3 higher vocational colleges in China for five consecutive years. In 2017, the college was awarded the title of China Higher Vocational College with top 50 international influencers. This research selected "Hotel English", as the test course of the course

examination, mainly because this course has obvious characteristics, integrating tourism English vocabulary, common sentence patterns and professional knowledge. It can fully demonstrate students' comprehensive abilities in listening, speaking, translation, communication, cooperation, service and emergency handling.

3.5. Data Collection

The course examination of this research was in early December 2021, which was consistent with the final examination. An anonymous questionnaire was distributed to all students after the exam. Overall satisfaction was measured using the scales of Wu (2020), Sha and Zheng (2021).

This scale measured overall test satisfaction from five aspects：

1) The examination can reflect the course content,

2) The examination content can promote the improvement of students' language ability,

3) The examination content was rich and reasonable,

4) The examination can be fair and trustworthy,

5) The examination can reflect the students' ability.

Students used a full-point scale to assess their overall satisfaction with the course exams. The five measurement indicators were equally weighted. The satisfaction ratio of each test group was calculated according to the weighted average method. A total of 90 questionnaires were distributed; 90 ones were received. The recovery rate was 100%

4. Results and Analysis

4.1. Compiling Factor Level Table

According to the experimental design, four factors were specified as the key independent variables. They were A, DL; B, NSA; C, PCG, and D, PPQ. The factor level table was compiled wherein the DL A has 3 levels of BA, OA and AA; the NSA, B has 3 levels indicated by 5%, 10% and 15%; the PCG, C has 3 levels indicated by 5%, 10% and 15% and the PPQ, D has 3 levels indicated by 5%, 10% and 15% respectively. The factor level table is shown in Table 1.

Table1. *Factor Level Table*

Factors / Levels	DL A	NSA B	PCG C	PPQ D
1	BA	5%	5%	5%
2	OA	10%	10%	10%
3	AA	15%	15%	15%

4.2. Orthogonal Test Results

As presented in (3.4) and (3.5), 90 students were divided into 9 groups and they took the "Hotel English" course examination at the same time. The satisfaction rate was tested. In order to reduce the error of the experiment, three teachers were assigned to be responsible for the scoring of NSA, CG and PPQ. The results thus obtained are tallied in Table 2.

Table2. *Orthogonal Test Results*

No.	Factor Level				Satisfaction Rate
	A	B	C	D	
1	1	1	1	1	0.72
2	1	2	3	2	0.74
3	1	3	2	3	0.79
4	2	1	3	3	0.75

5	2	2	2	1	0.72
6	2	3	1	2	0.82
7	3	1	2	2	0.69
8	3	2	1	3	0.72
9	3	3	3	1	0.71

4.3. Polar Variance Analysis

In the 1950s, the Japanese statistician, G. Taguchi, in order to use efficient experimental design to distribute points, proposed an orthogonal experimental design, to select the most representative experimental points, and to quickly find out, with fewer experiments, a combination of experimental factors for objective optimization (Yang, 2012). In this study, this method was used. The orthogonal table L9 (3^4) with 4 factors and 3 levels was selected, and the SPSSAU system was used to perform extreme variance analysis to obtain the K, Kavg and R, as shown in Table 3.

Table3. *Analysis of Orthogonal Test Results*

Index	**Level**	**Factor Level**			
		A	**B**	**C**	**D**
K	1	2.25	2.16	2.26	2.15
	2	2.29	2.18	2.20	2.25
	3	2.12	2.32	2.20	2.26
Kavg	1	0.75	0.72	0.75	0.72
	2	0.76	0.73	0.73	0.75
	3	0.71	0.77	0.73	0.75
R		0.06	0.05	0.02	0.04
Number of levels		3	3	3	3
repetitions per level		3	3	3	3
Factor primary and secondary relationship		A>B>D>C			
optimal level		2	3	1	3

As shown in the table above, the extreme variances R are 0.06 for the factor A, 0.05 for the factor B, 0.02 for the factor C, and 0.04 for the factor D. This means that the primary and secondary relationship of the factors goes: A>B>D>C. That is to say, the most important factor in the examination of tourism English major turned out to be Factor A, followed by Factor B, Factor D, and finally Factor C, the relatively the least influential factor. According to the size of K, the preferred solution is thus determined as OA of DL, 15% of NSA, 5% of PCG and 15% of PPQ. Through comprehensive analysis, the average value of students' test satisfaction is taken as the ordinate, and the level of each influencing factor is taken as the abscissa, as shown in Figure 1.

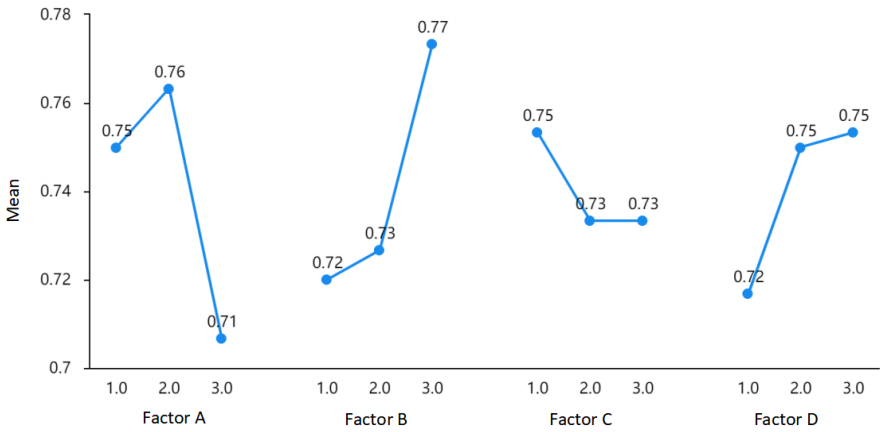


Figure1. *The level mean of the factors*

5. DISCUSSION

Through the above analysis of the results of the orthogonal test of the tourism English major in higher vocational college (TEMHVC), the following three aspects are discussed. The first was the influence

of various factors in the tourism English professional course examination. The second was the primary and secondary relationship of each factor. The third was the best combination of course exams.

5.1. The Influence of Various Factors

From the analysis results, it can be concluded that the influence of each factor on the course examination is different. Among them, DL, the difficulty level factor has the greatest impact on the course examination. But it is not the case that the lower the difficulty level of the course examination, the higher the students' satisfaction. What students want is a medium level of difficulty that matches their level of knowledge. In contrast, the factor of NSA, the proportion of non-standard answers, turned out to be the case: the higher the number, the stronger the satisfaction. Our interpretation is that the students prefer NSA in the course examination because such test items are more flexible and more autonomous. The impact of PCG, classroom grades on the course examination is not particularly large, neither the impact among 5%, 10%, and 15% is obvious. This is because the most important content of PCG is but daily attendance. Generally speaking, the attendance rate is good for almost all the students; therefore, PCG cannot discriminate the score differences, thus no significant impact whatever. As for the influence of the factors of PPQ, professional practice questions, students hope to have a certain proportion of PPQ. In this regard, the larger the proportion, the higher the satisfaction.

5.2. The Primary and Secondary Relationship of Each Factor

The results of extreme variance analysis show that the primary and secondary relationship of factors goes as what follows (4.3):

$$DL > NSA > PPP > PCG$$

It can be seen that DL, the difficulty level is the most concerned factor for students' test satisfaction, but it does not mean that the lower the difficulty, the higher the students' satisfaction as expected. Secondly, students expect more NSA, non-standard answer questions in the exam because NSA best reflects the flexibility and individual differences that must be taken into account in the curriculum exam. PPQ, the proportion of professional practice questions ranks third, indicating that students pay more attention to the practical use of English. The last factor is PCG, the proportion of classroom grades. As PCG is largely based on daily class attendance, students' satisfaction is basically in line with that of the course examination experience. In general, PCG is not high.

5.3. The Best Combination of Course Exams

Based on the analysis results of the orthogonal test, the optimal combination of the contents of the tourism English professional course examination is thus:

DL, the difficulty level is set on average (OA);

NSA, the non-standard answers account for 15%,

CG, the classroom grades account for 5%,

PPQ, the professional practice questions account for 15%.

Under the same learning conditions, just by appropriately adjusting the four factors: DL, NSA, PCG, and PPQ, in the content of the course examination, the satisfaction of students in the course examination can be well maintained. Also indirectly, the optimal combination reflects the degree of matching between students' satisfaction and test content. In this sense, the present research provides a good method to improve both the connotation and efficiency of students' course evaluation.

6. CONCLUSION

Three conclusions are drawn. First of all, the orthogonal experiment method can objectively reflect the influence of various factors, and get the best plan of the examination content. Such a method improved the matching degree between student satisfaction and students' mastery of course knowledge content. It increased the evaluation efficiency of course examinations. Next, through this research, the author found out the reasonable structure and proportion of the content of the tourism English course examination. It also shows us that the proportion of each influencing factor is not as high as possible. Instead, a reasonable percentage can produce better test results. Thirdly, the orthogonal experiment made possible the standardization of CETEM. As an achievement test, the

course exam for tourism English major often lacks uniform standards. Through case analysis, this study obtained the influence degree of each factor on the test, the optimal level of each factor and the best combination of test content. It is through the best combination that the standardized construction of course examinations can be achieved.

6.1. Significance

The course examination of higher vocational colleges is an important part of HVE, which affects the quality of training talents. This study can be used as a reference for foreign language curriculum in higher vocational colleges to increase the rationality of test content, improve evaluation and optimize test management, and testing methods. Students, as independent individuals, play not only the role of being tested, but also the role of evaluating the test. Foreign language majors are highly practical majors, and course examinations play an important role in screening talented personnel and in obtaining teaching feedback. The content of the exam is the chief influencing factor that students face directly; therefore, the proportion of related factors affects the satisfaction and quality of the entire test. At the same time, the rationality and scientificity of the enhanced course examination are obtained by using OTM. Students are the main body of teaching, and understanding their satisfaction with the course examinations is of great significance to the construction of foreign language majors and to the promotion of learning through evaluation.

6.2. Limitations

Three limitations remain. First, the sample size is small and exploratory factor analysis (EFA) did not conducted for questionnaire to obtain the justifiability of the four factors specified in the research. Secondly, from the perspective of language testing, the difficulty level was not calibrated using computer software for language testing. The difficulty level was subjectly presumed by researchers, instead. The third limitation concerns the research scope. As course examinations involves a wide range of stakeholders, the satisfaction should be studied more comprehensively.

6.3. Follow-up Studies

In view of the limitations inherent in this study, multiple sample sources and quantities can be used to analyze the influencing factors of course examinations in the future. As the influencing factors vary with different majors, the author is interested in establishing the multivariate influencing factor model in different language majors for further study. As there exist many stakeholders in course examination, the research object should be expanded to include employers, schools, parents and other groups as well as the results of the interaction. Hopefully, this would spark the interest of counterparts home and abroad.

7. ACKNOWLEDGMENTS

We would like to thank the guest editor of this special issue, Professor Zhang Quan as well as three anonymous reviewers for constructive comments and valuable guidance throughout the review process. We also thank Zhang Dianwei and Wu Jiao from the Vocational Research Center of Hebei Institute of International Business and Economics in China for their support and cooperation in this research.

REFERENCES

Chen, C. (2010) *Research on the evaluation of the quality of college entrance examination propositions from the* perspective *of customer satisfaction. Unpublished doctoral dissertation,* Tianjin University.

Gong, Y.Y. & Li, X.Y. (2004). Probability and Statistics. Wuhan: Huazhong University of Science and Technology Press.

Gui, S. C. (2007). *Gui Shichun's Anthology on English Language Education*. Beijing: Foreign Language Teaching Research Press

Gu, *Chinese Educational Dictionary: An updated and co-edited edition.* Shanghai: Shanghai Education M.Y. (1998). *Edu Press.*

Gui, S. C. (2017). *Selected Academic Papers of Gui Shichun.* Shanghai: Shanghai Foreign Language Education Press

Gui, S. C. (1986). *Standardized Testing - Theory, Principles and Methods*. Guangzhou: Guangdong Higher Education Press.

Guo, C. M. (2015). Discussion on the teaching reform of "Mechanical Manufacturing Technology" in higher vocational colleges. Economist. (1), 5-7.

Institute of Vocational and Technical Education Center of the Ministry of Education. (2016). China Vocational Education 2030 Research Report - Development Goals, Main Issues, Key Tasks and Promotion Strategies. *China Vocational and Technical Education* (25), 13.

Li, W. (2000).MET in China: *reform, explore and practice.* Higher Education Press.

Lin, Y. T., Chen, H. L. & Chen, J. S. (2018). *Exploring cognitive diagnostic transformation and in-depth analysis of language proficiency tests- Taking the Guangzhou English* Academic Test as an example Psychological Science, 41(4), 7-11.

Liu, D. H. (2012). *Research on the Theory and Method of Integrated Optimal Design* Based on Intelligent *Computing and Knowledge Guidance*. fense Industry Press.

Liu, K. (2019). *Research on vocational education of vocational students.* (Doctoral dissertation, Jiangxi Normal University).

Liu, Y. N. (2015). Research on the design of the test plan for elementary Chinese oral class. (*Doctoral dissertation, Hebei Normal University*).

Ma, S.C. (2010). Educational Services and the Sustainable Development of Higher Vocational Education. *Journal of Zhejiang Vocational and* Technical College of Industry and Trade, 10 (3), 1.

Rogers, C. R. (2015). On Human Growth. (Shi, M. L.). World Book Publishing Company Beijing Company

Rogers, C. R., & Freiberg, J. (2015). *Free Study (3rd edition).* (Wang, Y. H.). People Post Press.

Rowntree, D. (1988). *Dictionary of Western Education.* (Chen, J. P.). Shanghai Translation Publishing House.

Sheng, Q. X. (1997). *Ancient Chinese examination system*. Beijing: Commercial Press.

Wu, J. (2020). *Research on the reform of college course examinations from the perspective of humanism-taking N colleges as an example.* (Doctoral dissertation, Central China Normal University).

Wang, X. D. (2004). *A communicative test of college English reading comprehension.* (Doctoral dissertation, Southeast University).

Wang, X. Y., Zhong, Z. X., & Xiao, L. (2020). Exploration of the student-centered curriculum reform of "Computer Application Fundamentals"-Inspiration from innovative teaching methods based on the perspective of Fengjia University's teaching scene. *Yuzhang Teachers College Journal* (3), 4-7.

Wu, S. (2020). Listening to the voices of those involved in the test to build a satisfactory test for the people-A survey on the needs of the development of the national English proficiency test. *Foreign Languages* (2), 2-11.

Wu, S., & Zheng, H. S. (2021). College English CET-4 and CET-6 Satisfaction Survey-An empirical analysis based on structural equation modeling. Chinese Exam, (2020-4), 55-64.

Wu, X. M. (2011). Vocation-oriented application-oriented university course examination reform. *China Adult Education* (10), 3-5.

Xie, H. B. (2021). Research on the evaluation method of students' course performance based on the orthogonal test method. *Education in Heilongjiang-Higher Education Research and Evaluation* (8), 73-75.

Xu, Y. (2015). *Analysis of 333 Education Comprehensive Examination*. Beijing: Beijing Institute of Technology Press.

Yang, X. W. (2012). *Multi-objective drug synthesis optimization analysis of central experimental design based on micro-genetic algorithm.* (Doctoral dissertation, Shanxi Medical University).

Zhang, Q. (2016). (Ed.). Pacific-Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceeding .ISBN 978-981-10-8137-8 the Springer. doi: 10.1007/978-981-10-8138-5

Zhang, Q. & Yang, H. (2014). (Eds.). Pacific-Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceeding. ISBN 978-3-662-47489-1. ISBN 978-3-662-47489-1 (eBook) the Springer. doi: 10.1007/978-3-662-47490-7

Zhou, Y. (2015). An Analysis of the Teaching Method of Public Calligraphy for Non-normal Majors in Colleges and Universities. *New Curriculum Research*. Mid (9), 2-4.

Citation: *Liu Jiajie. "Towards Better Evaluation on Course Examination of Tourism English Major for Higher Vocational Education in China" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 15-24. DOI: https://doi.org/10.20431/2349-0381.09S1003*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 25-35
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1004
www.arcjournals.org

RASCH Model and Test Equating in China ---- A Comparison and Contrast of WINSTEPS and GITEST

Wu Jinyu

University of Electronic Science and Technology of China, Zhongshan Institute, China

***Corresponding Author:** **Wu Jinyu,** PhD, University of Electronic Science and Technology of China, Zhongshan Institute, China. Dr. Wu has been actively involved in language testing since her PhD studies in 2015. 12316603@ qq.com*

Abstract: *This paper addresses the basic concept, definition and practice regarding test equating. In China, test equating plays a central role for large-scale examinations with high stakes. It is still held as the prerequisite condition for item banking in computerized as well as in Internet-based testing. For the purpose of comparison and contrast, both Winsteps and GITEST are used to calibrate the same group of data of 40 Chinese students of non-English major collected from parallel tests through linking items ad hoc administered in a university in Guangdong Province, China. The item difficulty parameters thus obtained turned out to be 99.8% correlated. Comparison and contrast of these two types of software are elaborated. The paper concludes that Winsteps and GITEST are equally good for conducting test equating.*

Keywords: *Test equating, Rasch Model, CAT, Winsteps, GITEST, PROMS, RASCH-GZ*

1. Research Background

Over the past decades, especially since the Pacific-Rim Objective Measurement Symposium (PROMS) was first held in Malaysia in 2005, the academic community has paid increasingly attention to the application of Rasch model to the research of objective measurement. PROMS entered China for the first time in 2012. As the PROMS organizers in China, the authors believe that among various kinds of Rasch-based computer software available for test equating, both GITEST and Winsteps are great software programs to consider. They offer a wide range of application of Rasch model to practical testing problems, assumes binary scoring of item responses and gives stable and accurate estimates of item parameters and scale scores for both long and short tests and classroom exercises. They are best representing respectively the current applications based on the Rasch model in China and outside China. This paper attempts to present, among many of their features, the significant aspect of Winsteps: equating for parallel tests based on a group of minimum yet representative data and comparison and contrast with GITEST.

2. Test Equating and Its Significance

Though Winsteps is widely used for objective measurement of various purposes, test equating is seldom applied. Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004). Many testing programs use multiple forms of the same test. Such as college admission, in which serious decisions are made about people who might have taken the test at different administrations during a year or in different years, the primary reason for having multiple forms of a test is to maintain security and fairness. However, even though test developers attempt to construct test forms that are as similar as possible with each other in content and statistical specifications, the forms typically differ somewhat in difficulty. The comparability of tests scores across different tests measuring the same ability is an issue of considerable importance to test developers, measurement specialists, and test takers alike (Hambleton, Swaminathan & Rogers, 1991). Take the Matriculation English Test (MET) in China for example, which is the most prestigious and competitive examination of high stake administered annually to approximately 10 million candidates in China, and 60% or so of the participants can be enrolled.

Its item difficulties and test security must be put well under control and thus great importance is attached to it. If the same MET paper is administered repeatedly to different candidates nationwide annually to admit students for college studies, or if the same test paper is used repeatedly to different students before annual graduation for achievement evaluation, there is no way yet of protecting test security after its administration. On the other hand, it would not be feasible to administer two separate tests at once to the same group of candidates so as to compare the item difficulties of the tests. In this sense, equating plays a central role.

3. TEST EQUATING AND ITS CONCEPT

The concept of "equating" discussed in the present paper therefore refers to linking of separate 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 (Hambleton & Swaminathan, 1985, cited in Gui & Li, 1989). The idea is better illustrated as follows:

Group A test takers took Test A, which has L items with n anchor items;

Group B test takers took Test B, which has L items with n anchor items.

This is interpreted in language testing as two parallel test forms being written, each consisting of "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"s items of the two tests and put them on the same scale.(Zhang & Hu, 2000; Zhang, 2004). 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, 16 items in each test.

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 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 57 58 59 60 61 62 63

This is considered as the typical examples in terms of "equating of parallel tests". In today's testing practice, equating plays a central role and is held as the prerequisite condition for Computerized Adaptive Testing (CAT), item banking and for online testing in the forthcoming 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 adjusted. As equating is a complicated process requiring enormous data processing, and manual calculation is by no means feasible, Rasch-based computer software like Winsteps and GITEST offers us an effective tool. In what follows, we present a pair of representative yet real data to demonstrate the complete procedure of how equating is complemented by both GITEST and Winsteps. (Zhang & Hu, 2000; Zhang, 2004)

4. EQUATING BY GITEST: A SIMPLIFIED EXAMPLE

4.1. GITEST Program

GITEST is a Rasch-based system first developed by Ph.D program of applied linguistics headed by Prof. Gui Shichun(1986, 1990, 1993) of Guangzhou Institute of Foreign Languages, China as early as in 1980s. It was written in BASIC according to Rasch Model which is good at performing the following functions:

- It assumes binary (right-wrong) scoring;
- Designed for applications of both CTT and Rasch to practical testing problems;
- Maximum likelihood (ML);
- Tests of fit for individual items;
- Analysis of multiple subtests in one pass;
- Item analysis and test paper evaluation and report;
- Feedback for teaching and testing improvement ;

- Linking of 2 test forms through common items ;
- 200 items by10,000 candidates (maximum sample size) in a single run;

4.2. GITEST Data Editing

The data editing for GITEST is simple. The rows of data matrix are the test takers‟ ID followed by all the dichotomous responses presented by each test taker, while each column contains one answer to the corresponding test item. GITEST accepts two types of data responses: integer or char. Like all the other Rasch and IRT programs processing data of dichotomous in nature, if integer data are used, '1' represents right answer and „0‟, wrong answer. If char data are input, a line of key answers should be provided and put at the first line of the data matrix as shown in Table 1 and 2 in the following. Though written in Basic, GITEST can process the data metrics up to 10,000 persons by 200 items with a single run. This is the only Rasch-based software ever used to process data for Ten- Year Equating Project of Matriculation English Test (MET) funded by National Education examination Authority (NEEA) under Ministry of Education from 1990-1999 in China. (Gui, Li & Zhang, 1993)

Table1. *GITEST Integer Data Matrix*

TestA0001 11101011111111010101010100100000111111011

TestA0002 11110101011101010101010100101100111111101

TestA0003 111010111111110101010101001001111111111011

TestA0004 11110101010000010101111100100100111111101

TestA0005 11101011111111010111110100100000111111011

… … … … … …

TestA0079 00000001010101111101010100100010111111101

TestA0080 11101111111111010101010100111111111111011

Table2. *GITEST Char Data Matrix*

Key AABCBDCBDACDBADCBBCBCDABAABAAAACDBADCDAAAD

TestA0001 DCCCBDCBDACAAAACBBCBCDABAABACAAADBADCDAACC

TestA0002AACCDDCBDACDBAACDBCBCDABAABACAACDBADCDADCC

TestA0003ACCCBDCBDACDBAACBBCBCDABAABACAACDBADCDAADC

TestA0004ACCCBDCBDACDBAACBBCBCDABAABACAACDBADCDADCC

TestA0005ACDCBDDBDDCDBAACBDCBCDABDBBACBBBBBADCDABCC

… … … … …

TestA0079ABCCBDCBDACDBAACBBCBCDABAABACAACDBADCBAACC

TestA0080ACCCBDCBDACBBBACCCCBCDABAABACAACDBADCDAACC

4.3. GITEST Equating

With a single run, GITEST can yield the measure of the 16 linking items in both Test A and Test B thus obtained and listed in Table 3 below.

Table3. *GITEST: Linking Item Difficulties in logits of Test A and B*

ITEM	Test A	Test B
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	-0.027778	-0.608
0013	0.462	0.311
0014	0.213	-0.068
0015	-0.449	-0.189
0016	-0.555	-0.669
MEAN	-0.016	-0.181

What we are interested in here is the means of the 16 linking items in the two tests. As observed at the bottom of the table, the two means of the same linking items in Test A and Test B turned out to be different: -0.016 (logits) in Test A and -0.181 (logits) in Test B. The question is then raised: Why did the difficulties of the same 16 items turn out to be different?

Our interpretation is that the test items to which these common items are linked respectively in Test A and B are different. If -0.016 - (-0.181), the difference obtained from the mean minus the mean is 0.165 logit, indicating the test items in Test A are a little bit easier than those in Test B. 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 A are 0.165 easier in logit than those in Test B. "In such an example, the linking items are the hard items in **EASY** test but the easy items in the **HARD** test" (Wright & Stone, 1979; Zhang & Hu, 2000; Zhang, 2004). And the measure of the other items in both Test A and Test B obtained are listed in Table 4 below.

Table4. *GITEST Equated Item Difficulties*

ITEM	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
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	

5. EQUATING BY WINSTEPS: A SIMPLIFIED EXAMPLE

5.1. Winsteps Porgram

Winsteps is a Rasch-based program developed by John M. Linacre in 1984, which constructs Rasch measures from simple rectangular data sets, usually of persons and items. It is good at performing

more functions than GIETST does. For example, Winsteps can process up to 9,999,999 persons by 60,000 items with rating scales up to 255 categories for each item. At the same time, Winsteps provides a familiar "pull-down" user interface, intended to provide the user with maximum speed and flexibility. (Linacre, 2016)

5.2. Winsteps Data Matrix

To input data into Winsteps system for equating, we need create specific data matrix. We open an Excel spreadsheet, of which the first row is the variable names, each row is one person (subject, case), and each column contains one variable. Table 5 below briefly shows the idea.

Table5. *Excel spreadsheet used for Winsteps data matrix for Test Equating*

TestA0001 DCCC BDCBDACAAAACBBCBCDABAA BACAAADBADCDAACC
TestA0002 AACCDDCBDACDBAACDBCBCDABAA BACAACDBADCDADCC
TestA0003 ACCCBDCBDACDBAACBBCBCDABAA BACAACDBADCDAADC
TestA0004 ACCCBDCBDACDBAACBBCBCDABAA BACAACDBADCDADCC
TestA0005 ACDCBDDBDDCDBAACBDCBCDABDBBACBBBBBADCDABCC
… … … …
TestA0079 ACCCBDCBDACDBAACBBCBCDABAABACAACDBADCDADCC
TestA0080 ACDCBDDBDDCDBAACBDCBCDABDBBACBBBBBADCDABCC

TestB0001 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD
TestB0002 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD
TestB0003 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD
TestB0004 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD
TestB0005 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD
… … … …
TestB0072 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD
TestB0073 BACBBBBBADCDABCC ACBDCBAAABCDACBDCBCDCBACBD

5.3. Winstep Equating

Test Equating via linking items are straightforward with Winsteps, but do require prudent care. The more thought is put into test construction and data collection, the easier the equating will be. Such test equating proceed by Winsteps goes thus: Winsteps initially analyzes the linking items from the two tests, i.e. Test A and Test B and then analyzes each test separately. In Winsteps, the item parameter values can be anchored using command IAFILE=. Anchoring facilitates equating test forms and building item banks.

With this, a single run of Winsteps would let us obtain the item measures for all the items and construct the scale. In this step, separate analyses for each test were conducted with the 16 quality linking items anchored at the value that had been calibrated in the above step to the general item measures for all of the items. See Table 6 below.

TABLE6. *16 Linking Item STATISTICS: MEASURE ORDER*

2 CATS WINSTEPS 3.92.1

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%
10	24	153	1.55	.24	1.00	.0	.91	-.3	.35	.33	83.6	85.0
6	46	153	.59	.19	1.05	.6	1.05	.5	.31	.36	71.1	73.0
13	47	153	.56	.19	.94	-.7	.93	-.6	.42	.36	74.3	72.5
1	53	153	.35	.18	.97	-.4	.94	-.6	.40	.36	70.4	70.0
4	55	153	.28	.18	1.14	2.0	1.15	1.5	.21	.36	63.8	69.1
14	57	153	.22	.18	.94	-.9	.90	-1.0	.43	.36	71.1	68.4
5	58	153	.19	.18	1.11	1.6	1.19	2.0	.23	.36	65.1	68.1
8	65	153	-.03	.18	1.10	1.6	1.14	1.7	.25	.36	61.2	65.9
7	68	153	-.13	.17	1.05	.9	1.05	.7	.31	.36	60.5	65.5
15	71	153	-.22	.17	.84	-3.0	.79	-3.0	.53	.36	73.0	65.0
3	72	153	-.25	.17	1.08	1.4	1.08	1.0	.28	.36	64.5	64.9
11	75	153	-.34	.17	.95	-.9	.99	-.2	.40	.36	69.7	64.7
2	80	153	-.49	.17	.91	-1.6	.88	-1.6	.45	.35	66.4	64.5
16	81	153	-.52	.17	.91	-1.6	.87	-1.7	.45	.35	63.2	64.5
12	89	153	-.76	.18	.94	-1.0	.90	-1.2	.41	.34	67.8	65.6
9	97	153	-1.01	.18	1.09	1.4	1.13	1.3	.23	.33	61.8	67.8
MEAN	64.9	153.0	.00	.18	1.00	.0	.99	-.1			68.0	68.4

where the item difficulties were tailed in the order of decrease with Item 10, the hardest one (1.55 logits), for which merely 24 out of 135 test takers got the correct answer, and Item 9, the easiest one (-1.01 logits), for which 97 out of 135 got the correct answer, where TOTAL COUNT indicates totally 153 test takers from two groups taking respectively Test A and Test B tried these 16 linking items, and both INFIT and OUTFIT of the items were accepted. Table 7 below shows the linking item difficulties in logits of both Test A and B produced by Winsteps, indicating high correlationship with those of GITEST.

Similar to the previous steps, the first round of the analysis was undertaken to identify the under fit persons whose OUTFIT or INFIT MNSQ were larger than 2.0, and the second round of the analysis, without the under fit persons identified in the first round of the analysis, was used to calibrate the difficulty estimates for all of the items. In Winsteps, any items showing misfit to the Rasch model, i.e., the OUTFIT or INFIT MNSQ was larger than 2.0, were removed from the scale. No items were identified by this criterion and removed. Furthermore, any items with extremely high or low difficulty were investigated by experts specialized in English to determine whether they were appropriate for inclusion in the assessment. Consequently, no items were removed because their difficulties were all appropriate for the corresponding grades of the sampled test takers. The remaining items comprised the item pool of the two tests. (Linacre, 2016) The item measures for Test A and B of both pre-and post-equating are presented in the following Table 8, 9, 10 and 11 respectively.

Table7. *WINSTEPS: Linking Item Difficulties in logit of Test A and Test B*

ITEM	Test A	Test B
0001	0.33	0.06
0002	-0.24	-0.98
0003	-0.1	-0.7
0004	0.15	0.12
0005	-0.02	0.12
0006	0.15	0.74
0007	-0.07	-0.43
0008	-0.24	-0.07
0009	-0.98	-1.17
0010*	0.31	1.48
0011	-0.07	-0.85
0012	-1.04	-0.61
0013	0.46	0.31
0014	0.21	-0.07
0015	-0.45	-0.19
0016	-0.55	-0.67
MEAN	-0.13	-0.18
CORR with GITEST	0.8	1

Where all the values are observed highly correlated with those yielded via GITEST EXECPT Item 10 which in some way obviously affected the correlation.

Table8. *The Item Measures for Test A (Pre-Equating)*

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEASUR-AL		EXACT	MATCH	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	21	80	.53	.26	.96	-.3	.95	-.3	.28	.20	73.8	73.7	Q1
2	25	80	.27	.25	.98	-.2	.95	-.4	.27	.21	66.3	69.0	Q2
3	21	80	.53	.26	1.04	.4	1.03	.3	.12	.20	73.8	73.7	Q3
4	20	80	.60	.26	.99	.0	.98	-.1	.22	.20	75.0	75.0	Q4
5	35	80	-.29	.23	.95	-.8	.96	-.6	.31	.22	65.0	60.4	Q5
6	20	80	.60	.26	1.05	.5	1.07	.5	.08	.20	75.0	75.0	Q6
7	34	80	-.24	.23	1.04	.7	1.03	.5	.15	.22	58.8	61.0	Q7
8	38	80	-.45	.23	.99	-.1	.99	-.1	.24	.22	56.3	59.3	Q8
9	19	80	.67	.27	1.05	.4	1.08	.6	.08	.19	76.3	76.2	Q9
10	31	80	-.07	.24	1.06	.8	1.06	.7	.10	.22	58.8	63.2	Q10

11	56	80	-1.44	.25	.97	-.3	.95	-.3	.28	.21	72.5	70.2	Q11
12	47	80	-.93	.23	.99	-.1	1.02	.2	.22	.22	67.5	61.7	Q12
13	26	80	.21	.24	1.13	1.4	1.16	1.5	-.06	.21	62.5	67.9	Q13
14	35	80	-.29	.23	.98	-.3	.97	-.4	.27	.22	57.5	60.4	Q14
15	20	80	.60	.26	1.00	.1	1.01	.1	.18	.20	75.0	75.0	Q15
16	20	80	.60	.26	1.00	.0	1.00	.0	.20	.20	75.0	75.0	Q16
17	30	80	-.02	.24	.99	-.1	.99	-.1	.24	.22	62.5	64.1	Q17
18	36	80	-.34	.23	.97	-.5	.96	-.6	.29	.22	61.3	59.9	Q18
19	24	80	.33	.25	.95	-.5	.93	-.6	.33	.21	67.5	70.1	Q19
20	27	80	.15	.24	.98	-.2	.98	-.2	.25	.21	67.5	66.9	Q20
21	40	80	-.55	.23	.97	-.6	.96	-.7	.29	.23	57.5	59.1	Q21
22	31	80	-.07	.24	1.00	.0	1.02	.3	.21	.22	66.3	63.2	Q22
23	16	80	.89	.28	.98	-.1	.92	-.4	.25	.18	80.0	80.0	Q23
24	28	80	.10	.24	1.02	.3	1.03	.3	.17	.22	61.3	65.9	Q24
25	50	80	-1.09	.24	1.01	.1	.99	-.1	.21	.22	58.8	64.0	Q25
26	15	80	.98	.29	.97	-.1	.90	-.4	.27	.18	81.3	81.2	Q26
27	24	80	.33	.25	1.00	.0	.99	.0	.21	.21	70.0	70.1	Q27
28	34	80	-.24	.23	.98	-.3	.99	-.1	.26	.22	63.8	61.0	Q28
29	31	80	-.07	.24	1.06	.8	1.05	.6	.11	.22	61.3	63.2	Q29
30	27	80	.15	.24	1.07	.8	1.10	1.0	.06	.21	67.5	66.9	Q30
31	30	80	-.02	.24	1.07	1.0	1.09	1.1	.06	.22	60.0	64.1	Q31
32	27	80	.15	.24	1.08	1.0	1.07	.7	.05	.21	62.5	66.9	Q32
33	31	80	-.07	.24	.93	-1.0	.92	-1.0	.36	.22	66.3	63.2	Q33
34	34	80	-.24	.23	1.08	1.4	1.12	1.7	.04	.22	56.3	61.0	Q34
35	48	80	-.98	.23	1.03	.4	1.03	.5	.16	.22	58.8	62.4	Q35
36	13	80	1.16	.31	.96	-.1	.92	-.2	.26	.17	83.8	83.7	Q36
37	31	80	-.07	.24	.94	-.9	.95	-.5	.34	.22	71.3	63.2	Q37
38	49	80	-1.04	.24	.97	-.4	.97	-.4	.28	.22	70.0	63.2	Q38
39	22	80	.46	.26	.95	-.4	.92	-.5	.31	.20	71.3	72.5	Q39
40	26	80	.21	.24	.98	-.2	.97	-.2	.25	.21	70.0	67.9	Q40
41	38	80	-.45	.23	.94	-1.1	.94	-1.0	.34	.22	63.8	59.3	Q41
42	40	80	-.55	.23	.94	-1.3	.92	-1.5	.37	.23	57.5	59.1	Q42
MEAN	30.2	80.0	.00	.25	1.00	.0	1.00	.0			66.8	67.1	
P.SD	9.9	.0	.57	.02	0.05	.6	.06	.6			7.2	6.5	

Table9. *The Item Measures for Test B (Pre-Equating)*

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEASUR-AL		EXACT	MATCH	
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	Item
1	29	73	.06	.25	1.03	.4	1.04	.4	.24	.29	57.5	64.1	Q27
2	46	73	-.98	.25	.84	-1.9	.81	-1.7	.51	.27	69.9	65.7	Q28
3	41	73	-.67	.25	.95	-.7	.93	-.7	.36	.28	69.9	63.1	Q29
4	28	73	.12	.25	1.13	1.5	1.15	1.4	.10	.29	58.9	64.8	Q30
5	28	73	.12	.25	1.09	1.0	1.10	1.0	.16	.29	64.4	64.8	Q31
6	19	73	.74	.28	.97	-.1	.93	-.3	.33	.28	74.0	74.9	Q32
7	37	73	-.43	.24	.98	-.3	.96	-.4	.32	.29	58.9	62.0	Q33
8	31	73	-.07	.25	1.05	.8	1.07	.8	.20	.29	57.5	62.9	Q34
9	49	73	-1.17	.26	.90	-1.0	.88	-.8	.41	.26	71.2	68.6	Q35
10	11	73	1.48	.34	.96	-.1	.97	.0	.29	.25	86.3	85.5	Q36
11	44	73	-.85	.25	.92	-1.0	.89	-1.0	.40	.28	67.1	64.4	Q37
12	40	73	-.61	.25	.93	-1.1	.90	-1.1	.40	.28	65.8	62.7	Q38
13	25	73	.31	.26	.98	-.2	.98	-.1	.31	.29	68.5	67.8	Q39
14	31	73	-.07	.25	.97	-.4	.97	-.3	.33	.29	71.2	62.9	Q40
15	33	73	-.19	.25	.89	-1.7	.86	-1.6	.46	.29	65.8	62.2	Q41
16	41	73	-.67	.25	.88	-1.7	.86	-1.5	.46	.28	72.6	63.1	Q42
17	24	73	.38	.26	.88	-1.2	.83	-1.4	.47	.28	71.2	68.9	Q43
18	20	73	.66	.27	1.07	.6	1.11	.7	.17	.28	74.0	73.7	Q44
19	36	73	-.37	.24	1.14	2.1	1.14	1.6	.09	.29	49.3	61.9	Q45
20	17	73	.90	.29	1.07	.5	1.06	.4	.17	.27	78.1	77.6	Q46
21	39	73	-.55	.25	1.11	1.6	1.13	1.4	.12	.28	58.9	62.4	Q47

22	41	73	-.67	.25	1.11	1.6	1.12	1.2	.12	.28	56.2	63.1	Q48
23	43	73	-.79	.25	1.05	.7	1.34	2.9	.14	.28	64.4	63.9	Q49
24	16	73	.98	.29	1.05	.4	.98	.0	.22	.27	76.7	78.9	Q50
25	28	73	.12	.25	.92	-.9	.88	-1.1	.41	.29	64.4	64.8	Q51
26	13	73	1.26	.32	1.14	.7	1.12	.5	.08	.26	80.8	82.9	Q52
27	38	73	-.49	.24	1.02	.4	1.01	.2	.25	.29	60.3	62.2	Q53
28	35	73	-.31	.24	1.00	.1	.99	-.1	.29	.29	61.6	61.9	Q54
29	35	73	-.31	.24	.92	-1.2	.89	-1.2	.41	.29	64.4	61.9	Q55
30	29	73	.06	.25	1.02	.3	1.02	.2	.26	.29	60.3	64.1	Q56
31	22	73	.52	.27	.91	-.8	.87	-.8	.42	.28	75.3	71.3	Q57
32	35	73	-.31	.24	.95	-.8	.92	-.9	.37	.29	67.1	61.9	Q58
33	38	73	-.49	.24	1.01	.1	.99	-.1	.28	.29	54.8	62.2	Q59
34	15	73	1.07	.30	.89	-.6	.88	-.5	.42	.27	80.8	80.2	Q60
35	28	73	.12	.25	1.03	.4	1.07	.7	.23	.29	61.6	64.8	Q61
36	20	73	.66	.27	1.13	1.0	1.20	1.2	.07	.28	68.5	73.7	Q62
37	29	73	.06	.25	1.07	.9	1.09	.9	.18	.29	60.3	64.1	Q63
38	44	73	-.85	.25	1.05	.7	1.06	.6	.20	.28	58.9	64.4	Q64
39	43	73	-.79	.25	.96	-.5	.93	-.6	.35	.28	67.1	63.9	Q65
40	15	73	1.07	.30	1.04	.3	1.25	1.1	.14	.27	80.8	80.2	Q66
41	16	73	.98	.29	.96	-.2	.91	-.4	.34	.27	79.5	78.9	Q67
MEAN	30.5	73.0	.00	.26	1.00	.0	1.00	.0			67.2	67.6	
P.SD	10.2	.0	.68	.02	.08	1.0	.12	1.0			8.3	6.8	

Table10. *The Item Measures for Test A (Post-Equating)*

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEASUR-AL		EXACT	MATCH	DISPLACE	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP%	OBS%	EXP%		
36	13	80	1.55A	.34	1.21	.9	1.17	.7	.26	.15	83.8	87.6	-.32	Q36
26	15	80	1.05	.29	.97	-.1	.90	-.4	.27	.18	81.3	81.3	.00	Q26
23	16	80	.97	.28	.98	-.1	.92	-.4	.25	.18	80.0	80.0	.00	Q23
9	19	80	.74	.27	1.05	.4	1.09	.6	.08	.19	76.3	76.3	.00	Q9
4	20	80	.67	.26	.99	.0	.98	.0	.22	.20	75.0	75.0	.00	Q4
6	20	80	.67	.26	1.06	.5	1.07	.5	.08	.20	75.0	75.0	.00	Q6
15	20	80	.67	.26	1.01	.1	1.01	.1	.18	.20	75.0	75.0	.00	Q15
16	20	80	.67	.26	1.00	.1	1.00	.0	.20	.20	75.0	75.0	.00	Q16
1	21	80	.60	.26	.96	-.2	.95	-.3	.28	.20	75.0	73.8	.00	Q1
3	21	80	.60	.26	1.04	.4	1.04	.3	.12	.20	72.5	73.8	.00	Q3
32	27	80	.59A	.26	1.26	2.0	1.27	1.8	.05	.20	65.0	73.6	-.36	Q32
39	22	80	.56A	.26	.97	-.3	.93	-.4	.31	.20	71.3	73.0	-.03	Q39
19	24	80	.41	.25	.95	-.5	.93	-.6	.33	.21	67.5	70.2	.00	Q19
27	24	80	.35A	.25	.98	-.2	.97	-.2	.21	.21	72.5	69.2	.06	Q27
2	25	80	.35	.25	.98	-.2	.95	-.4	.27	.21	66.3	69.1	.00	Q2
13	26	80	.29	.24	1.13	1.4	1.16	1.5	-.06	.21	62.5	68.0	.00	Q13
30	27	80	.28A	.24	1.08	.9	1.12	1.1	.06	.21	67.5	67.9	-.05	Q30
20	27	80	.23	.24	.98	-.2	.98	-.2	.25	.21	67.5	67.0	.00	Q20
40	26	80	.22A	.24	.96	-.4	.96	-.4	.25	.21	71.3	66.9	.06	Q40
31	30	80	.19A	.24	1.11	1.3	1.15	1.5	.06	.22	61.3	66.4	-.14	Q31
24	28	80	.17	.24	1.02	.3	1.03	.3	.17	.22	61.3	66.0	.00	Q24
17	30	80	.06	.24	.99	-.1	.99	-.1	.24	.22	62.5	64.2	.00	Q17
10	31	80	.00	.24	1.06	.9	1.06	.7	.10	.22	58.8	63.3	.00	Q10
22	31	80	.00	.24	1.00	.0	1.02	.3	.21	.22	66.3	63.3	.00	Q22
34	34	80	-.03A	.23	1.11	1.6	1.16	2.0	.04	.22	58.8	62.8	-.14	Q34
33	31	80	-.13A	.23	.91	-1.4	.91	-1.4	.36	.22	67.5	61.4	.13	Q33
7	34	80	-.16	.23	1.04	.7	1.03	.5	.15	.22	58.8	61.1	.00	Q7
5	35	80	-.22	.23	.95	-.8	.96	-.6	.32	.22	65.0	60.5	.00	Q5
14	35	80	-.22	.23	.98	-.3	.98	-.4	.27	.22	57.5	60.5	.00	Q14
41	38	80	-.22A	.23	.96	-.7	.96	-.6	.34	.22	58.8	60.5	-.16	Q41
29	31	80	-.25A	.23	1.03	.6	1.02	.4	.11	.23	55.0	60.2	.25	Q29
18	36	80	-.27	.23	.97	-.5	.97	-.6	.29	.23	61.3	60.0	.00	Q18
37	31	80	-.34A	.23	.92	-1.5	.93	-1.3	.34	.23	65.0	59.4	.34	Q37
8	38	80	-.38	.23	.99	-.1	.99	-.1	.24	.23	56.3	59.3	.00	Q8
21	40	80	-.48	.23	.97	-.6	.96	-.7	.29	.23	57.5	59.2	.00	Q21
28	34	80	-.49A	.23	.99	-.3	.99	-.1	.26	.23	62.5	59.2	.32	Q28
42	40	80	-.52A	.23	.94	-1.3	.92	-1.4	.37	.23	57.5	59.2	.04	Q42
38	49	80	-.76A	.23	.95	-.9	.95	-.8	.28	.22	61.3	60.7	-.21	Q38
12	47	80	-.85	.23	.99	-.1	1.02	.3	.22	.22	67.5	61.7	.00	Q12

35	48	80	-1.01A	.24	1.05	.7	1.06	.7	.16	.22	58.8	63.9	.10	Q35
25	50	80	-1.02	.24	1.01	.1	.99	-.1	.21	.22	58.8	64.0	.00	Q25
11	56	80	-1.37	.25	.97	-.3	.95	-.3	.28	.21	72.5	70.2	.00	Q11
MEAN	30.2	80.0	.07	.25	1.01	.0	1.01	.0			64.4	67.2	.00	
P.SD	939	.0	.60	.02	.07	.8	.08	.8			7.4	7.0	.12	

Table11. *The Item Measures for Test B (Post-Equating)*

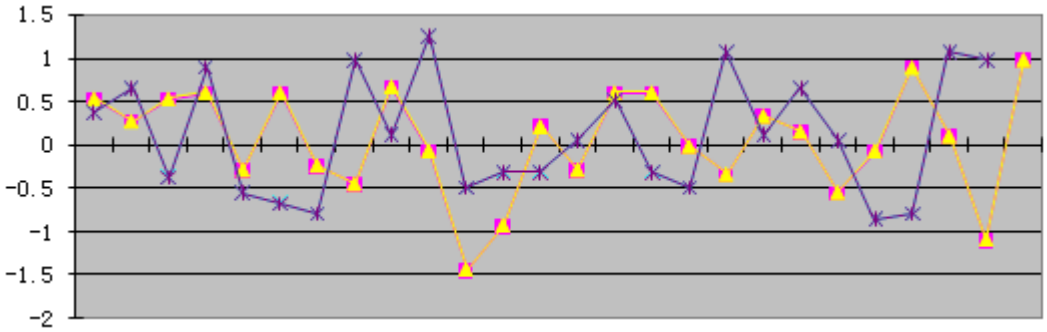
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP%	EXACT OBS%	MATCH EXP%	DISPLACE	Item
10	11	73	1.55A	.33	.90	-.4	.90	-.3	.29	.25	86.3	84.2	.11	Q36
26	13	73	1.44	.32	1.13	.7	1.11	.5	.08	.26	80.8	82.8	.00	Q52
34	15	73	1.25	.30	.88	-.6	.88	-.5	.42	.26	80.8	80.2	.00	Q60
40	15	73	1.25	.30	1.04	.3	1.24	1.1	.14	.26	80.8	80.2	.00	Q66
24	16	73	1.16	.29	1.05	.4	.97	-.1	.22	.27	76.7	78.9	.00	Q50
41	16	73	1.16	.29	.96	-.2	.91	-.4	.34	.27	79.5	78.9	.00	Q67
20	17	73	1.08	.29	1.07	.5	1.06	.4	.17	.27	78.1	77.5	.00	Q46
18	20	73	.84	.27	1.07	.5	1.11	.7	.17	.28	74.0	73.7	.00	Q44
36	20	73	.84	.27	1.13	1.0	1.19	1.2	.07	.28	68.5	73.7	.00	Q62
31	22	73	.70	.27	.91	-.8	.87	-.9	.42	.28	75.3	71.2	.00	Q57
6	19	73	.59A	.26	.88	-1.2	.84	-1.2	.33	.28	75.3	69.4	.33	Q32
13	25	73	.56A	.26	1.00	.0	1.01	.1	.31	.28	69.9	68.9	-.07	Q39
17	24	73	.56	.26	.88	-1.2	.83	-1.4	.47	.28	71.2	68.9	.00	Q43
1	29	73	.35A	.25	1.05	.7	1.07	.6	.24	.29	58.9	65.4	-.11	Q27
25	28	73	.30	.25	.92	-.9	.88	-1.1	.41	.29	64.4	64.8	.00	Q51
35	28	73	.30	.25	1.03	.4	1.07	.7	.23	.29	61.6	64.8	.00	Q61
4	28	73	.28A	.25	1.12	1.5	1.14	1.4	.10	.29	58.9	64.6	.02	Q30
30	29	73	.24	.25	1.02	.3	1.02	.2	.26	.29	60.3	64.0	.00	Q56
37	29	73	.24	.25	1.07	.9	1.09	.9	.18	.29	60.3	64.0	.00	Q63
14	31	73	.22A	.25	.99	-.1	.99	-.1	.33	.29	68.5	63.9	-.10	Q40
5	28	73	.19A	.25	1.07	.9	1.08	.8	.16	.29	67.1	63.6	.11	Q31
8	31	73	-.03A	.25	1.05	.7	1.05	.6	.20	.29	57.5	62.0	.14	Q34
28	35	73	-.13	.24	1.00	.0	.98	-.1	.29	.29	61.6	61.9	.00	Q54
29	35	73	-.13	.24	.92	-1.2	.89	-1.2	.41	.29	64.4	61.9	.00	Q55
32	35	73	-.13	.24	.95	-.8	.92	-.9	.37	.29	67.1	61.9	.00	Q58
7	37	73	-.13A	.24	.98	-.2	.96	-.4	.32	.29	61.6	61.9	-.12	Q33
19	36	73	-.19	.24	1.13	2.0	1.14	1.6	.09	.29	49.3	61.8	.00	Q45
15	33	73	-.22A	.24	.89	-1.8	.86	-1.6	.46	.29	67.1	61.8	.21	Q41
3	41	73	-.25A	.24	.95	-.8	.93	-.8	.36	.28	64.4	61.9	-.24	Q29
27	38	73	-.31	.24	1.02	.4	1.01	.2	.25	.28	60.3	62.1	.00	Q53
33	38	73	-.31	.24	1.01	.1	.99	-.1	.28	.28	54.8	62.1	.00	Q59
11	44	73	-.34A	.24	.91	-1.4	.90	-1.2	.39	.28	68.5	62.2	-.33	Q37
21	39	73	-.37	.24	1.11	1.6	1.13	1.4	.12	.28	58.9	62.4	.00	Q47
22	41	73	-.49	.25	1.11	1.6	1.12	1.2	.12	.28	56.2	63.0	.00	Q48
2	46	73	-.49A	.25	.82	-2.8	.79	-2.3	.51	.28	79.5	63.0	-.31	Q28
16	41	73	-.52A	.25	.89	-1.7	.86	-1.4	.46	.28	72.6	63.2	.03	Q42
23	43	73	-.61	.25	1.05	.7	1.33	2.9	.14	.28	64.4	63.8	.00	Q49
39	43	73	-.61	.25	.96	-.6	.93	-.6	.35	.28	67.1	63.8	.00	Q65
38	44	73	-.67	.25	1.05	.7	1.06	.6	.20	.27	58.9	64.3	.00	Q64
12	40	73	-.76A	.25	1.00	.0	.98	-.1	.40	.27	64.4	65.3	.33	Q38
9	49	73	-1.01A	.26	.91	-.9	.88	-.8	.41	.26	71.2	68.8	.02	Q35
MEAN	30.5	73.0	.18	.26	1.00	.0	1.00	.0			67.5	67.4	.00	
P.SD	10.2	.0	.66	.02	.08	1.0	.12	1.0			8.4	6.7	.12	

In Table 10 and Table 11 above, items asterisked with „A" indicating they are Anchored Items, i.e. used as linking items between the two tests; therefore, all the test items whose difficulties are rescaled in the similar fashion as discussed in 3.3 above and are comparable on the same scale. This shows us that the equating results obtained from GITEST and Winsteps are the same: of the two tests, Test A is easier as can be observed in Table 12 below. And a careful examination of the parameters obtained further reinforces the assumption proposed by Wright & Stone (1979), i.e. "the linking items are the hard items in EASY test but the easy items in the HARD test". This also shows us that these two types of software are much of the same in terms of equating and are genuinely Rasch-based.

Table12. *Comparison of Equated Test Items Produced by GITEST and Winsteps.*

Item	TestA-GITEST	TestA - Winsteps	TestB -GITEST	Test B- Winsteps
0017	0.528	0.06	0.378	0.38
0018	0.273	0.27	0.661	0.66
0019	0.528	0.53	-0.369	-0.37
0020	0.596	0.6	0.896	0.9
0021	-0.29	-0.29	-0.548	-0.55
0022	0.596	0.6	-0.669	-0.67
0023	-0.237	0.97	-0.791	-0.79
0024	-0.449	-0.45	0.98	0.98
0025	0.667	0.67	0.118	0.12
0026	-0.073	1.05	1.258	1.26
0027	-1.445	-1.44	-0.488	-0.49
0028	-0.927	-0.93	-0.309	-0.31
0029	0.213	0.21	-0.309	-0.31
0030	-0.29	-0.29	0.055	0.06
0031	0.596	0.6	0.516	0.52
0032	0.596	0.6	-0.309	-0.31
0033	-0.018	-0.02	-0.488	-0.49
0034	-0.344	-0.34	1.068	1.07
0035	0.335	0.33	0.118	0.12
0036	0.154	0.155	0.661	0.66
0037	-0.555	-0.55	0.055	0.06
0038	-0.073	-0.07	-0.852	-0.85
0039	0.895	0.89	-0.791	-0.79
0040	0.096	0.1	1.068	1.07
0041	-1.092	-1.09	0.98	0.98
0042	0.98	0.977		

Corr:
TEST A: 0.9855
TEST B: 0.9999



FigureI. *Item Difficulties of both Test A and Test B obtained from GITEST and Winsteps*

6. CONCLUSION AND DEVELOPMENT

From the above analyses and discussion, we could come to the conclusion that Winsteps and GITEST are different but alike and their properties can be summarized as what follows:

At the first glance, Winsteps and GITEST seem so different because data matrix for GITEST to process is simply a smaller text file, while for Winsteps, an Excel worksheet doc is needed and the data matrix that can be processed is much bigger. What's more, WINSTEPS can perform more statistical analyses and plotting. In contrast, GITEST handles classic test analyses and Rasch only. If Winsteps is international and paid to use it, GITEST is local but free. Yet, these two types of software, apparently different in some way, possess an affinity with each other. They are both Rasch-based and work well for test equating via anchored or linking items. Both are capable of reporting error messages. On the whole, GITEST and WINSTEPS: each has its own merits and one cannot be replaced by the other. Their utility largely depends on the user's need and purpose.

It is because of these reasons that efforts have been ad hoc made during this COVID-19 pandemic period since 2019 to have successfully updated GITEST which can run online, process sample size of unlimited number of items by unlimited number of subjects, produce desired plotting, testing reports and provide online technical support as WINSTEPS does.

REFERENCES

Benjamin, D. Wright & Mark H. Stone. (1979). Best test design Rasch measurement. MESA.

Gui, S.C. Li, W. & Zhang,Q. (1993). The Application of IRT to Test Equating for MET in China. PP.39`1-393. In NEEA. (Ed.), the 4th China National Education Test Conference Proceeding, China Peace Press. ISBN 7-80101-089-2/G.64

Gui,S.C.(1986). Standardized examination: theory, practice and method. Guangdong Higher Education Press.

Hambleton, R., Swaminathan, H., Rogers, H. (1991). Fundamentals of item response theory. Sage Publications, Inc.

Kolen, M. J., & Brennan, R.L. (2004). Test equating, scaling and linking. Springer Verlag New York, Inc. ISBN: 9780387400860

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

Mok, M.M.C & Zhang, Q. (2018).Introduction to Rasch measurement [Translation]. JAM Press. ISBN 978-1-934116-13-5

Mok, M.M.C & Zhang, Q. (2015). Constructing Variables. Book of Abstracts Vol.I. Journal of Applied Measurement. ISBN 978-1-934116-11-1

Mok, M.M.C. & Zhang, Q. (2014). Constructing Variables. Book of Abstracts Vol.II. Journal of Applied Measurement. ISBN 978-934116-10-4

Zhang, Q. et al (2014). A Rasch-based approach for comparison of English listening comprehension between CET and GEPT. PP.115-130. In Zhang, Q. & Yang, H. (Eds.). Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings. Springer. ISBN 978-3-662-47489-1.

doi: 10.1007/978-3-662-47490-7.

Zhang, Q. (2004a). BILOG and PARSCALE: different but alike. In Banerjee, J & Wall, D. (Eds.). Language Testing Update (LTU) Issue 35- Spring 2004. Centre for Research in Language Education. Lancaster University Lancaster, England. International Language Testing Association Official Newsletter.

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

Zhang, Q & Hu, X.X. (2000). PC-BILOG and its Application in China: Equating and Item-banking. In Jose Lai & Tam, Pauline Po-yee. (Eds.) Crosslink in English Language Teaching. Volume 1, 2000. English Language Teaching Unit, the Chinese University of Hong Kong. ISSN 1605-7511

Citation: *Wu Jinyu. "RASCH Model and Test Equating in China ---- A Comparison and Contrast of WINSTEPS and GITEST" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 25-35. DOI: https://doi.org/10.20431/2349- 0381.09S1004.*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 36-45
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1005
www.arcjournals.org

RASCH-GZ: The 1st Chinese Version of RASCH-Based Item Analysis and Test Equating System (Part I)

Wei Jin-gang

Guangzhou Quantong Scientific Technology for Education Co. Ltd, Guangzhou, China.

Corresponding Author:** ***Wei Jin-gang, *PhD, Computer Engineer of Guangzhou Quantong Scientific Technology for Education Co. Ltd, Guangzhou, China. With years' of working experience, Dr. Wei has been actively involved in software application and development for language testing on large scale using java, python, html, delphi. weijingang@rasch-gz.com*

1. Research Background

It is known to all, since the last century, of all the objective measurement methods, researchers and practitioners have been using various software systems based on the Rasch model, such as: WINSTEPS, FACET, etc., because the Rasch model is the most versatile, powerful, and the most feasible!

In China, GITEST system can be deemed as one of such powerful and feasible software members. GITEST was originally developed as early as in 1980's under the guidance of the late famous Chinese linguist Prof. Gui Shichun (1933-2017). The system was programmed in BASIC, running under DOS processing data matrix of (200 items X 10,000 candidates, the maximum). From the viewpoint of Rasch professionals, GITEST is typically Rasch model-based item analysis and test equating system. It is worth mentioning that it is GITEST that was used to conduct the ten-year (1990-1999) equating project of Matriculation English Test (MET) sponsored by the National Education Examination Authority (NEEA) under Ministry of Education, China. MET is the most influential from a professional point of view; therefore, in this sense, GITEST has played a pivotal role in MET equating and has made a solid foundation for the standardized tests to be successfully implemented in China.

With the advent of the Internet era and the sustained 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[1] during the global fight against the COVID-19.

Technically, RASCH-GZ focuses on two main functions of language testing: item analysis and test equating. This chapter introduces the item analysis part of RASCH-GZ, and the application of test equating will be introduced in the next issue.

2. Comparison of GITEST and RASCH-GZ

The comparison between the fully upgraded RASCH-GZ and the GITEST system is shown in Table 1 below.

Table1. *Comparison of GITEST and 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**

[1]http://www.rasch-gz.com

3. ITEM ANALYSES

Rasch-GZ has two important functions: item analyses and test equating. Item analyses are based on classical test theory (CTT). Rasch-GZ helps analyze each option of MC questions and generates a test report based on the following parameter. The present section deals with analyses, and test equating will be presented with data and discussion in the next issue.

Table2. *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

Item analysis provides feedback about each option of a MC question and generates an analysis report for the entire test paper based on the information above so that teachers and test item writers could use the information above to moderate/improve the quality of their test item production or to adjust the item difficulties, or decide whether or not to abandon some poorly designed test items.

4. BASIC OPERATION OF RASCH-GZ SYSTEM

This section guides users to learn to use RASCH-GZ with the simple and easy-to-understand language, diagrams and texts and avoids complicated technical terms. The basic operation of RASCH-GZ is divided into six parts. In what follows will be introduced three parts: (4.1) booting and system login, (4.2) data input; (4.3) item analyses and the evaluation report for a whole test paper.

4.1. System Login

First, click on the professional version program icon to start and run the RASCH-GZ system. The following login interface appears on your computer screen.

Figure4.1. *Login interface of Rasch-GZ*

System login is performed as follows:

(1) Enter the registered user name;

(2) Enter the password;

(3) If you need help, click the "Help button" to display a demonstration of the help;

(4) To exit the system, click the "Exit" button;

(5) Click the "Login" button to enter the system, and the following interface will pop up on the screen, namely: the data entry interface of Rasch-GZ as shown in Figure 4.2 below.

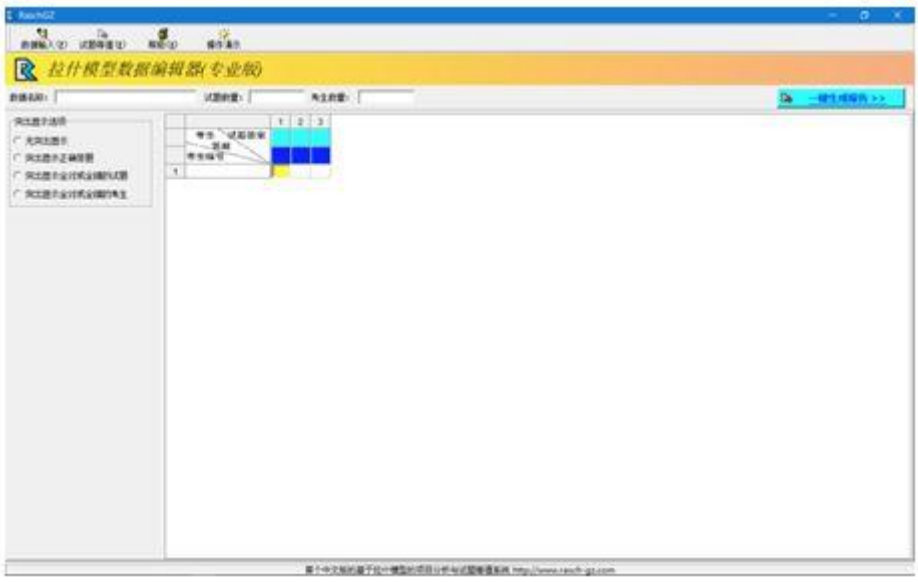


Figure4.2. *The data entry interface of Rasch-GZ*

4.2. Data input Method and Data File Storage

Like any other data processing system today, RASCH-GZ data entry is divided into manual entry and data file import. The data format accepts the text file format of GITEST and the data file format of EXCEL; the data type can be a char, such as: ABCDACDABACD, or an integer, such as: 101010101111100.

4.2.1. Manual Creation of Data Files

Manual creation of data files requires the following steps:

Step 1: Click "Data Input" in the upper left menu bar of Figure 4.2. The following interface will pop up, namely: the manual data input interface of the Rasch-GZ system as shown in Figure 4.3 below.

Figure4.3. *The manual data input interface of the Rasch-GZ system*

Step 2: Click the "Add New Test Items Section" button in Figure 4.3 to create the data file you want. At this time, we get the following interface, that is, the establishment of the data file format setting interface as shown in Figure 4.4 below.

Figure4.4. *The manual data file format setting interface*

Step 3: Please enter your own defined file name, number of test items, number of test takers and length of test takers' ID number according to the prompts in Figure 4.4 above. For example, in the first line, create your own first data file name in the space of the data name: GD08;

On the second line, enter 85 in the number of test items, indicating that we have 85 items; enter 100 in the number of test takers, indicating that we have 100 test takers; enter 9 in the box indicating the length of test takers' ID, indicating that the length of test takers' ID is 9 digits;

Starting from the third line, we entered each part of the test paper according to the prompts.

In the column of the serial number, we entered 1, 2, and 3 respectively, indicating that this test has three parts and the corresponding names of each part. They are LST (listening), GRM (grammar), and RDG (reading);

In the space of the number column, we entered the number of test items corresponding to the test, 20, 35, and 30 respectively, indicating that the listening part of this test has 20 items, the grammar part has 35 items, and 30 items in the reading section.

If the verification is correct, we will press the "OK" button to go to the next step.

Step 4: Click the "OK" button to create a data file according to the current data format, as shown in Figure 4.5 below:

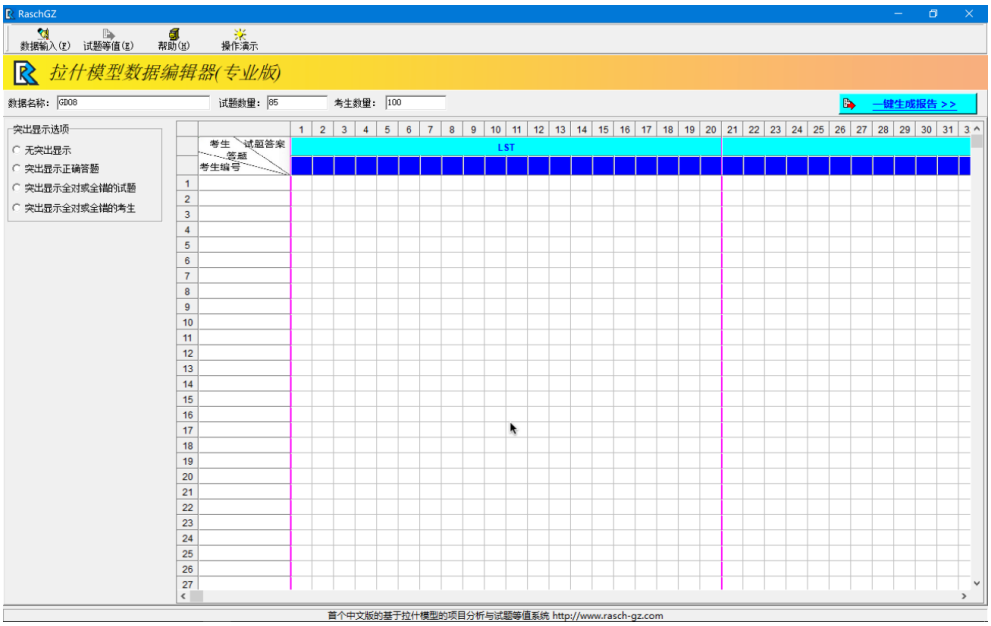


Figure4.5. *The data file format for manual inputting interface*

Step 5: now we start to manually input the data.

Enter the test takers' ID number in the form: Click on the first cell under "Student ID No." in the upper left corner of the form, a yellow input box will be displayed, and start to enter the test taker's ID number. The test takers' ID may be integers, or letters or combination of both. After the input is completed, press the Enter key to automatically enter the input box of the next test taker's ID until all test takers' IDs are input, as shown in Figure 4.6 below.

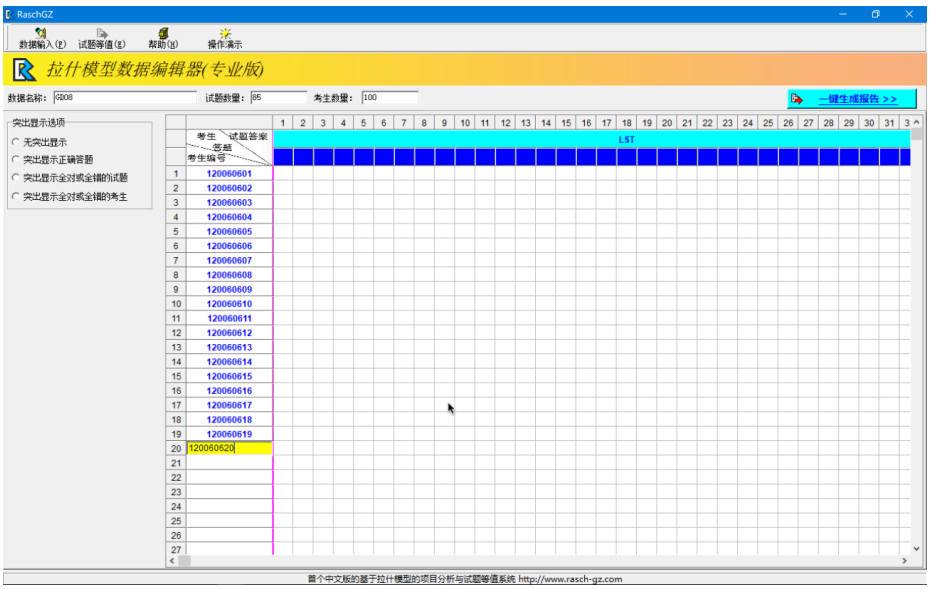


Figure4.6. *Rasch-GZ manual input test takers' ID data format interface*

Step 6: Enter the correct answers in the form:

Click on the first box with white background under “Examinees data” in the upper left corner of the form, a yellow input box will be displayed. Now, start to enter the correct answers of the test, i.e. keys, usually the characters A, B, C, and D. While entering the correct answers, enter one, press the Enter key once (yellow in the space behind), the system will automatically display next input box until all the answers have been entered.

In case the raw data are integers like 10101010001, it indicates that 1 means correct answer, and 0 means wrong answer. In such a case, no correct answers (keys) are provided, as shown in Figure 4.7 below

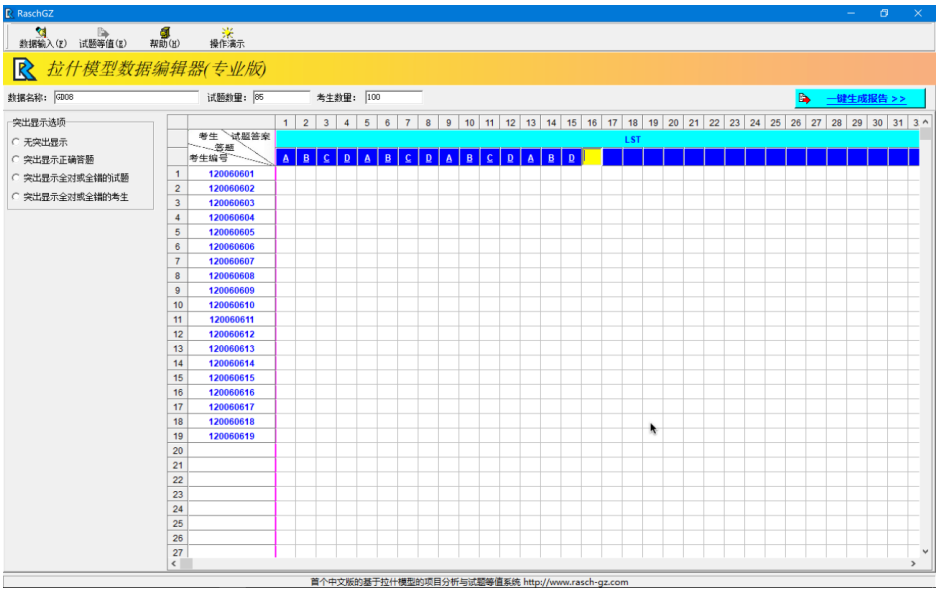


Figure4.7. *The interface for raw data of integers*

Step 7: Enter the raw data. Click the first cell with blue background on the right side of the "Raw Data" in the upper left corner of the table, a yellow input box will be displayed, and start to input the raw data. Usually, the original answers are Char such as A, B, C, or D, which are the options for multiple-choice question test type. When inputting these raw data, enter one char, press the Enter key once (the space behind it will appear yellow), the system will automatically display next box to input until all the raw data have been entered. Figure 4.8 below shows the basic data matrix format of the data file wherein the first column on the left is the test taker’s ID number, the first row above is the correct answer. Staring from the second row, we have the test takers’ raw data.

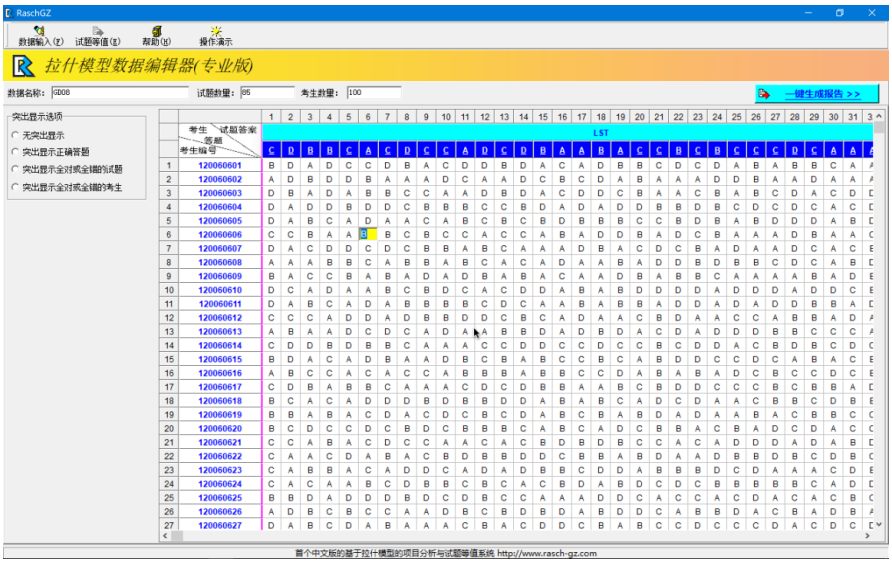


Figure4.8. *Basic data matrix format of RASCH-GZ data file*

Step 8: Save the current data file

Once the raw data input is done, check it! And make sure all are correct. From the "Data Input" menu in the upper left menu bar, click "Save" to save the currently input data. By this time, the system will pop up a window, indicating that you have successfully saved the data with the file name of "GD08" defined by yourself. At this time, the system automatically adds the suffix to "GD08.dat"

Step 9: One-click to quickly obtain the required analysis report file.

4.2.2. External Data File Import Method

If you already have a data file edited in Rasch-GZ format (see Figure 4.8 for details), you can directly import it into the Rasch-GZ system, and then with a click of mouse, you would quickly obtain an analysis report file. Currently Rasch-GZ accepts GITEST data text file format and EXCEL data file format.

(1) The way to import GITEST data edited in Rasch-GZ format goes as follows: Click "Import GITETEST Data" from the "Data Input" menu in the upper left corner of the menu bar, the system will pop up the following dialog box (see Figure 4.9), select the corresponding GETTEST format data file and open to the current data window to import data.

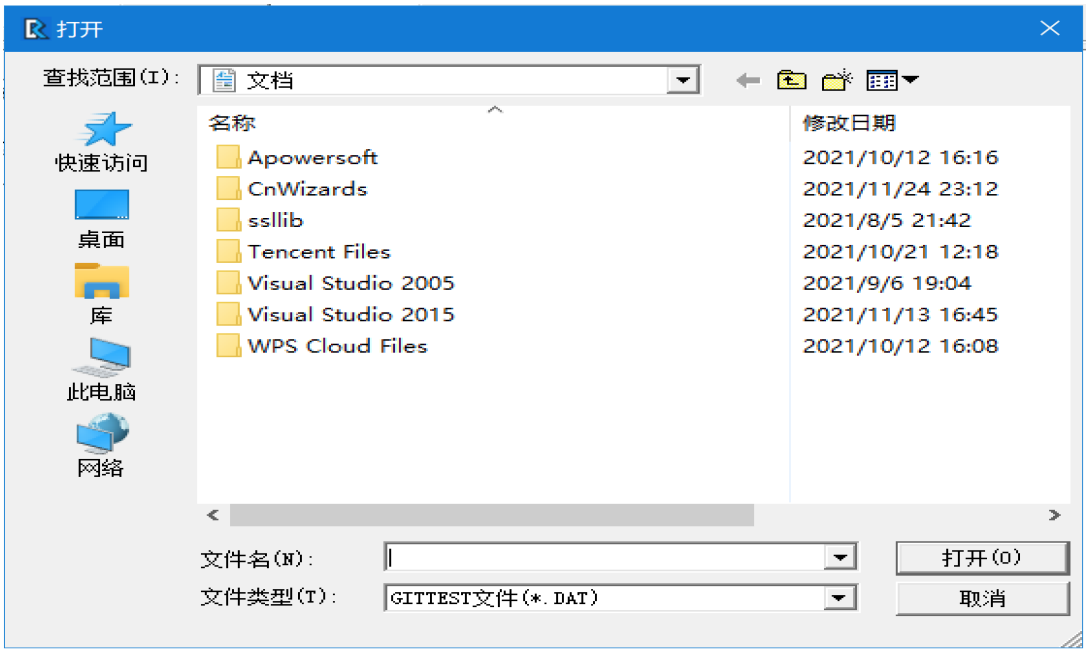


Figure4.9. *GITEST external data file import dialog*

(2) The way to import EXCEL data edited in Rasch-GZ format is as follows: Click "Import EXCEL Data" from the "Data Input" menu in the menu bar, the system will pop up the following dialog box (see Figure 4.10 below), select the corresponding EXCEL format data file, and open to the current data editing window to import data. Excel version supports older versions as well as versions after 2007. For details, see Figure 4.10 EXCEL External Data File Import Dialog Box.

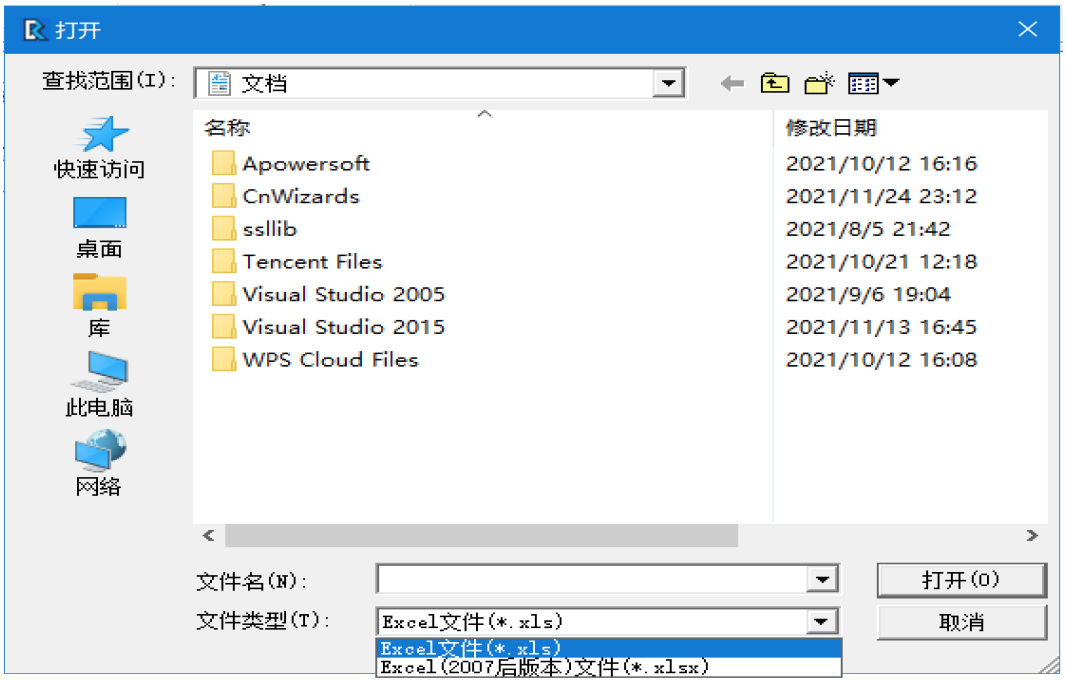


Figure4.10. *EXCEL external data file import dialog*

4.3. Obtain Data Analysis Reports

So long as the data file is imported successfully, or manually input is completed, and the user makes sure that all are correct, the required data analysis reports can be obtained. RASCH-GZ provides two ways to obtain data report files: the first one is to quickly generate all data analysis reports with one click; the second one is to click on the data analysis report specified by the user.

Click the "One-click Generate Analysis Report" button from the top of the table column. Rasch-GZ will automatically analyze the current data and generate relevant analysis reports within one or two seconds after the mouse click. If the data matrix is large, there will prompt a progressing bar during the data processing time. After the progress reaches 100%, we'll have the data analysis report viewing interface as shown in Figure 4.11.

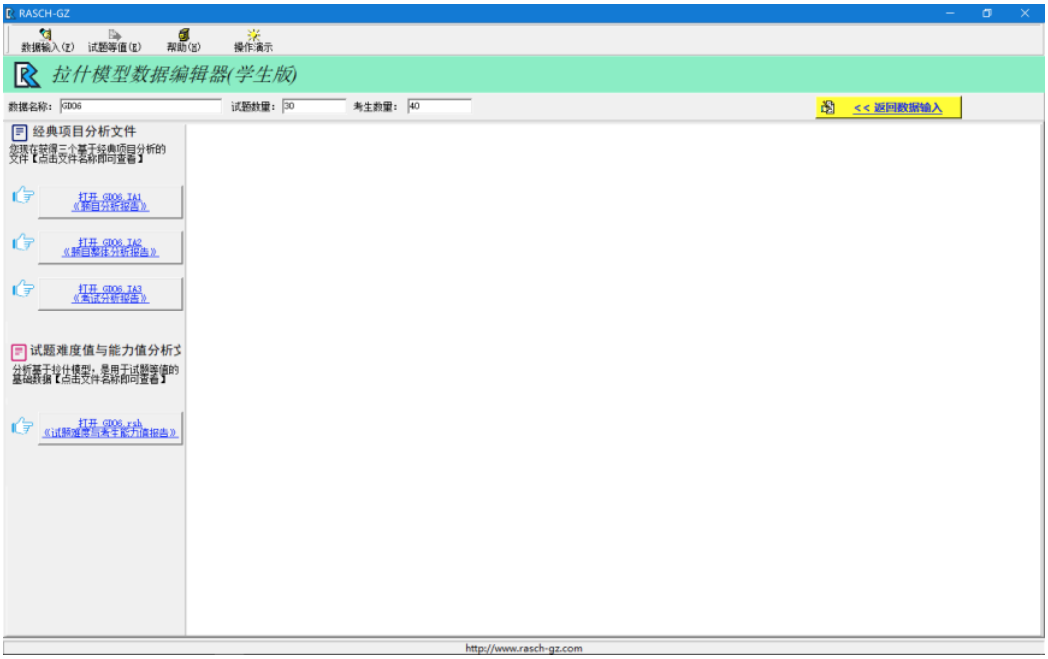


Figure4.11. *RASCH-GZ data analysis report viewing interface*

The interface description shown in Figure 4.11 is as follows:

There are four file names in the left column of Figure 4.11 above, with different suffixes. They are:

MFR Data 001.ia1 "Item Analysis Report for Each test item", (suffixed with ia1)

MFR Data 001.ia2 "Two-way report about difficulty-discrimination of all the test items in a whole test paper", (suffixed with ia2)

MFR Data 001.ia3 "An analysis report about a whole test paper ", (suffixed with ia3).

These three files, (as mentioned above), are based on the classic test theory, with the suffixes -.ia1-ia2 and –ia3 respectively.

Apart from this, there is another file with the suffix **–rsh,** i.e. MFR data 001.rsh.This file is the "test item difficulty/test taker ability report" based on Rasch model. **This data file with –rsh ending is an indispensable file for test equating.**

Sometimes, users only need to obtain a specific data file according to their own research purpose. In such a case, just click the required data analysis report. See Figure 4.12 below for details. Users need to get a specific data file interface.

Figure4.12. *Users' specific data file interface*

4.4. Other Data Related Operations

4.4.1. How to Open a Saved Data File

Click "Open" from the "Data Input" menu in the menu bar, and the "Open Data File" window will pop up. See Figure 4.13 below for details.

4.4.2. How to Open Recently Used Data Files

Click "Open Recently Used" from the "Data Input" menu in the menu bar, and the file names of the 5 recently used data files would be displayed in the lower menu. Click the selected file name to open the data and use it.

4.4.3. How to Moderate the Property Value of the Current Data

Click "Properties" from the "Data Input" menu in the menu bar to pop up the "Data Properties" window. Modifiable values include: data name, number of test items, number of test takers, length of test takers' ID, name and number of each section of the test paper.

Enter the new corresponding attribute value in the corresponding input box, and you can add or delete operations in the test section. Click the "OK" button to apply the modified properties, click the "Cancel" button to discard the modification.

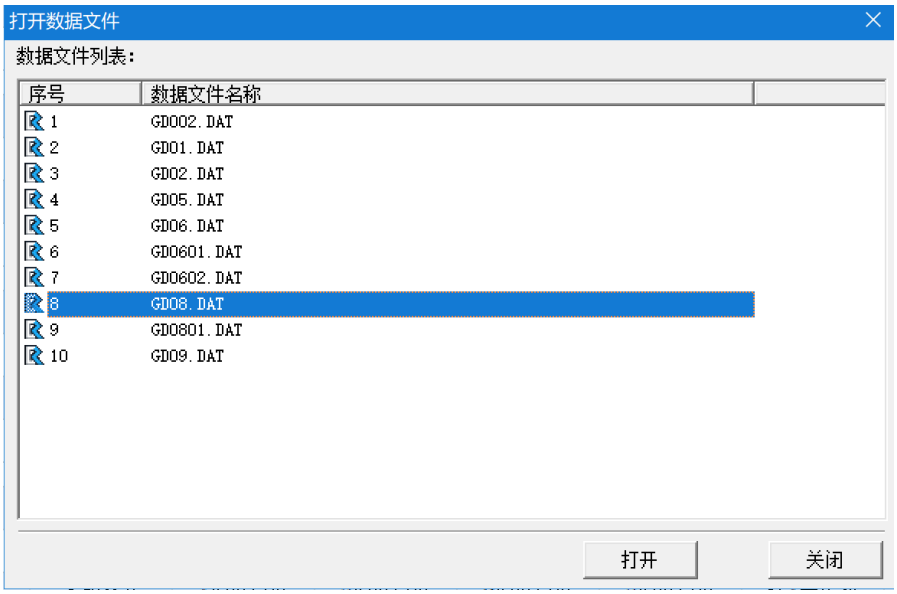


Figure2.13. *Rasch-GZ "Open Data File" Interface*

4.4.4. Close the Current Data

Click "Close" from the "Data Input" menu in the menu bar to close the data currently being edited.

4.4.5. Save the Current Data as

Click "Save As" from the "Data Input" menu in the menu bar, and the "Data Properties" setting window will pop up to modify the new data name. Save the data currently being edited to a new data file.

5. SUMMARY

The Rasch model can provide a complete solution to almost every measurement problem encountered in the social sciences, and is especially suitable for researchers in professional fields such as language testing because the raw data of such disciplines are difficult to control and the concept is vague. The most updated Chinese version of RASCH-GZ has 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 www.rasch-gz.com. The student version comes with a small data matrix (30 items x 40 candidates) and all 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. If the result is satisfactory, but the amount of data to be analyzed is large, and test equating is needed, then the users could apply for the professional version of Rasch-GZ online. (To be continued)

Citation: *Jin-gang Wei. "RASCH-GZ: The 1st Chinese Version of RASCH-Based Item Analysis and Test Equating System (Part I)" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 36-45. DOI: https://doi.org/10.20431/2349-0381.09S1005*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 46-58
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1006
www.arcjournals.org

Text Complexity of Reading Comprehension Passages in National Matriculation English Test: A Three-level Corpus Study

Yin Kailan

China

Corresponding Author:** ***Yin Kailan, *Exams, British Council, Chongqing, China. She actively engages in National Social Science Foundation Program and also is experienced in organizing large scale exams like IELTS and Aptis. Kailan.Yin@outlook.com*

Abstract: *This study investigated the text complexity of reading comprehension passages in China's National Matriculation English Test (NMET) of year 2020 and 2021, in the purpose of providing validation evidence for new NMET reform. Text complexity of 76 reading passages has been measured and compared on the three dimensions: lexical level, syntactic level, and discourse level. The natural language processing tools used in the study included Coh-Metrix and Eng-Editor. T-test and Wilcox test were conducted to compare the difference of each indicator.*

The results suggested that the lexical level text complexity revealed the most evident changes between the two years. Significant elevation was found in lexical diversity of the NMET reading passages, in which the lexical diversity of 2021 NMET reading passages increased moderately compared with that of year 2020. The syntactic level text complexity also showed an inflation in noun phases density in 2021 compared to that of 2020. Of the discourse level text complexity, insignificant increase of the indices occurred throughout the two years and the general trend was not necessarily rising. Nevertheless, the decrease of average hypernymy for verbs gave evidence of the growing text abstractness of NMET reading passages in 2021. Combined, the results might indicate that text complexity of the reading passages in the NMET from 2020 to 2021 has been steadily increasing by including low frequency and academic vocabulary, diversifying vocabulary in the passages, and complicating sentence structures. The results were further examined against the New English Curriculum Standards and guidelines to analyze whether the changes were reflected in the policies. It unraveled that the exams required a much larger vocabulary size than the number indicated in the guidelines, and more often, of thematic context and genre, the passages of the two years' NMET employed unproportioned use of human and society and exposition. Suggestions for test designers and pedagogical practices were provided accordingly.

Keywords: *National Matriculation English Test, text complexity, reading comprehension, Coh-Matrix, corpus-based study*

1. INTRODUCTION

It has been 45 years since the resumption of Gaokao in 1977. Over the past years, great changes have taken place in all areas of China, with the economy advancing by leaps and bounds, culture becoming highly prosperous, and people's lives becoming increasingly affluent. This has laid a solid foundation for developing education in China, whereas also places higher demand on it. Within this frame of reference, the National College Entrance Examination, also known as Gaokao, a compelling examination that enrolls millions high school graduates, is being reformed constantly in accordance with the education requirements (Liu, 2017, 2019). In regards of English compulsory subject of Gaokao, the National Matriculation English Test (NMET) has also been taking advantage of language testing research, and considering the needs of enrollment system as well as the high school English teaching and learning, making timely adjustment to the content and form of NMET.

It is known to all, standardized examination was first introduced into China by the late Chinese famous linguist Professor Gui Shichun (1930-2017) who is also the first professional conducting successfully the ten-year (1990-1999) Matriculation English Test (MET) equating project sponsored by Ministry of Education of China (Gui, 1985, 2007, 2017). From 1978 to 1988, The National

Education Examinations Authority (NEEA) began to adopt one unified English test paper for Gaokao. Within a decade, the English subject of Gaokao varied its test structure every year, designed many word and sentence level questions and emphasized on examining the candidates' English Language knowledge rather than their ability to use it (Liu, 2017). In 1985, the Ministry of Education decided to conduct a standardized reform trial of Gaokao in Guangdong Province, initiating the standardized examinations in China. The forerunners Prof. Gui Shichun and Prof. Li Xiaoju contributed a lot to the standardization of English subject of Gaokao. For instance, they proposed five procedures to measure and maintain test standardization, which includes observing score distribution, item analysis, improving rating reliability, normalizing scaled scores and test quating (Li et al., 1989). The English subject of Gaokao was required to develope in a standardized way ever since, and was called Matriculation English Test (MET). In this period, the MET instruction was published by NEEA, multiple-choice questions began to dominate in MET, language skills and language use became the focus, and writing was introduced in MET (Qi, 2007). In 1991, the name National Matriculation English Test (NMET) was first employed and used to date. From 1991 to 1999, NMET inleted error correction, spelling and competing the dialogues, while reduced the number of grammar and vocabulary multiple choice questions. In 2000, NMET revolutionarily included listening in most provinces (Lv, 2017). In 2014, NEEA further adjusted the structure of NMET, replacing the previous single- sentence language use questions with discourse grammar fill-in-the-blank questions and launched a pilot in Zhejiang and Shanghai, whose NMET should include continuation writing and be conducted twice a year in 2016 (*A Year Two Test* reform). In 2021, Guangdong province, Jiangsu provinces and other 12 provinces/municipality adopted the *A Year Two Tests* mode.

During these reforms, reading comprehension still holds a dominating position in NMET, accounting for approximately 25% to 35% of the total NMET scores. The general NMET guideline made by NEEA requires the candidates to understand common topics, illustrate the main idea, structure and details, deduct the meaning of specific words and phrases and finally understand the opinions, purpose and attitude of the passages (Wang, 2018). However, few studies have investigated the textual characteristics of the passages after the *New NMET Reform*, whose new mode was employed by 14 provinces in 2021. Therefore, facilitated by the natural language processing tools, this study aims to examine the text complexity of the NMET reading comprehension passages from 2020 to 2021, in the purpose of providing validation evidence for New NMET Reform. It probes into the real condition in the text selection and presentation of the NMET reading test, and offers suggestions for the test developing and pedagogically activities.

2. Literature Review

2.1. Reading Comprehension and Text Complexity: Definition and Development

Successful comprehension of written reading assessment tasks is influenced by a variety of factors, such as the test taker's cognitive ability, knowledge, and motivation,; or test task characteristics such as task description, wording and format of questions, and context (Kintsch & Kintsch, 2005). In general, these factors can be grouped into three categories: reader, task, and text. Understanding the factors that influence reading comprehension can provide test researchers and educators with a deeper understanding of test develop in a variety of academic areas, including reading, science, and social studies (Khalifa & Weir, 2009). In this study, the author puts a spotlight on the textual factors that influence reading comprehension that is text complexity.

Text complexity usually refers to the difficulty of a text and, in a narrow sense, equals to the linguistic features that affect text comprehension (Guo, et al., 2018). Research has shown that when text complexity is similar to the language level of foreign language learners, it helps to develop learners' language competency (Crossley et. al, 2012), while when text difficulty is much higher or lower than learners' language proficiency, it may hinder learners' language development (Kontovourki, 2012). Therefore, it is important to select text material of appropriate difficulty for learners.

Research on text complexity can be used to guide the design of reading tasks, daily classroom assessments, and to assess students' language proficiency in large-scale examinations (Lyashevskaya et.al, 2021). Previously, researches have measured text complexity with readability formulas such as the Flesch Readability the Fresch-Kincaid Grade Levels. Such formulas are easy to manipulate and

can visually detect text readability, but they examine the superficial text characteristics of the passages, and ignore syntactic, articulatory and semantic factors (Kintsch & Kintsch, 2005). The development of natural language processing and computational linguistics has prompted text complexity studies to incorporate deeper text features such as semantics, rhetoric, coherence, etc. (Guo, et al., 2018), typically represented by tools such as Coh-Metrix, Reading Maturity Metrix, and Text Evaluator. Jin et al. (2018) designed an Eng-Editor, a tool that could be used to evaluate and adopt texts based on the proficiency level specified in China's Standards of English Language Ability (CSE). This is the first tool that originated in Chinese English learning and teaching context, whose corpus is composed of The New English Curriculum Standards for Compulsory Education, The English Curriculum Standards for General High Schools (hereinafter called The New English Curriculum Standards), and the past NMET tests etc.

2.2. Researches on Text Complexity of NMET Reading Comprehension Passages

Part of the domestic studies on NMET reading comprehension focus on the reliability and validity of the content of reading comprehension questions in the college entrance examination co- temporally or over time (Gu, & Wang, 2008; Tao, 2017); the morph symbol ratio, lexical density, syntactic difficulty, and text length of NMET reading comprehension passages (Hu, 2018; Chen & Zhang, 2020). Most of these articles refer *The New English Curriculum Standards* and the syllabus for NMET. By Comparing the NMET reading comprehension with the two documents mentioned above from different perspectives, the researchers explore whether the development of NMET reading comprehension meets the requirements of the latter, or whether it is consistent with the documents, so as to judge the content validity of the tests (Xiao, 2014). However, the exploration of text complexity is not yet comprehensive. For example, studies mostly use the Flesch Readability of reading formula to calculate ease of reading, and the index has certain shortcomings. On the other hand, studies either start from vocabulary or syntax to explore text difficulty, and fewer combine the three aspects of vocabulary, syntax and discourse to explore comprehensively. Even fewer studies (Huang & Wang, 2020) have explored the text complexity of reading comprehension passages after the new NMET reform.

2.3. Lexical Level, Syntactic Level and Discourse Level Text Complexity

Word count and word length are the most direct measures of lexical level text complexity, which means that the more the number of longer words, the more difficult the text is to read, and studies have shown that it takes more time to process a longer word than a shorter word in English (Perfetti, 2011). It is also widely accepted in reading studies that readers with a large vocabulary could better understand the texts (RAND Reading Study Group, 2002). Inferring the meaning of a large number of new words based on a particular context will dampen learners' learning confidence (Far, 2016). Traditionally, lexical diversity has been calculated by type-token ratio (TTR). This formula with a larger type-token ratio indicating a more diverse vocabulary. Compared with TTR, Measure of Textual Lexical Diversity (MTLD) and VOCD, another two indices that could reflect the lexical diversity, were also less affected by the length of text (McCarthy & Jarvis, 2010). Lexical density is also an indicator of lexical complexity (RAND Reading Study Group, 2002).

Reading tests are often time-limited. Green et al. (2008) found that when time is limited, test takers become more stressed and their cognitive load increased accordingly. In general, longer sentences require more information to be processed and the accuracy of sentence comprehension is reduced (Far, 2016). Longer sentences and texts may also affect test takers' performance by reducing their working memory efficiency (Crossley et al., 2014). McNamara et al. (2014) argued that the shorter the sentence, the fewer words before the main verb or the fewer words before the noun phrase, the easier the syntax of the sentences is in the text; at the same time, readers may find the text more difficult when the density of passive voice sentences and negative sentences is too high. Latent semantic analysis (LSA) (Deerwester et al, 1990) and adjacent argument and stem overlap evaluation are effective way to decode the coherence of texts. LSA is "a mathematical method for computer modeling and simulation of the meaning of words and passages by analysis of representative corpora of natural text" (Landauer & Dumais, 2008). To construct a semantic space for a language, LSA first casts a large representative text corpus into a rectangular matrix of words by coherent passages, each cell containing a transform of the number of times that a given word appears in a given passage. The

matrix is then decomposed in such a way that every passage is represented as a vector whose value is the sum of vectors standing for its component words. Similarities between words and words, passages and words, and of passages to passages are then computed as dot products, cosines or other vector standing for its component words. Similarities between words and words, passages and words, and of passages to passages are then computed as dot products, cosines or other vector-algebraic metrics. Similarly, the Eng-Editor syntactic difficulty level will be encompassed to comprehensively examine the syntactic level text complexity of NMET reading passages.

Of discourse level text complexity, thematic contexts and genres, cohesion, text abstractness and readability are the four factors that influence the passages in this study. Specifically, all language learning activities should take place within a certain thematic context, for instance, and the New English Curriculum Standards provides 32 sub-themes based on different types of discourse and students should learn around such specific thematic context. Empirical studies have also shown that familiarity with a topic or genre influences test takers' performance on high-stakes tests (Crossley et al., 2012). While it is clear from the above discussion that word and syntax have an impact on reading, while the impact of the cohesion on reading remains controversial (Green et al., 2010). Cohesion refers to the specific elements of a text that indicates the coherent feature of the text and facilitates readers' comprehension (McNamara et al., 2014). A better understanding of the importance of cohesion in comprehension was the main inspiration for Crossley et al. to develop Coh-Metrix ("Coh" in Coh-Metrix means cohesion) (Crossley et al., 2014).

It is suggested that abstract text is more difficult to understand than content words or images in many researches (Corkill et al., 1988), possibly because when processing abstract text, readers' cognition is confined to a single language system (verbal or nonverbal), whereas when dealing with concrete language, readers can draw on knowledge of both linguistic and non-linguistic systems (concrete language may be pictorial) to aid comprehension (Green et. al, 2010).

The popular understanding of text complexity is approximately equivalent to readability or easebility. Readability formulas thus will be essential indicators in this study, and use Flesch Readability, Flesch-Kincaid Grade Level, and Coh-Metrix L2 Readability, together with Eng-Editor difficulty level to measure the discourse level text complexity of the NMET reading comprehension passages. The indices that will be used in this study are listed in Table 1 below.

Table1. *The adopted framework of text complexity*

Lexical level text complexity (L)
L1 Word count
L2 Average word length
L3 Average word frequency for content words
L4 Type-token ratio
L5 MTLD
L6 VOCD
L7 Word beyond NMET syllabus
L8 Eng-Editor lexical difficulty level
Syntactic level text complexity (S)
S1 Average sentence length
S2 Noun phrase density
S3 Average modifiers per noun phrase
S4 Average words before main verb
S5 Agentless passive voice density
S6 Negation density
S7 Average argument overlap for adjacent sentences
S8 Average stem overlap for adjacent sentences
S9 Average LSA overlap for adjacent sentences
S10 Average LSA overlap for adjacent paragraphs
S11 Eng-Editor syntactic difficulty level
Discourse level text complexity (D)
D-TT Thematic context
D-G Genres

D-C1	Causal connectives incidence
D-C2	Logical connectives incidence
D-C3	Adversative and contrastive connectives incidence
D-C4	Temporal connectives incidence
D-A1	Average concreteness for content word
D-A2	Average hypernymy for nouns
D-A3	Average hypernymy for verbs
D-R1	Flesch Readability
D-R2	Flesch-Kincaid Grade Level
D-R3	Coh-Metrix L2 Readability
D-R4	Eng-Editor text difficulty level

3. METHOD

This section briefly introduces the research questions, research materials, instruments, and research procedure for analyzing the data.

3.1. Research Questions

1) What are the similarities and differences in the results of the lexical level text complexity of the NMET reading comprehension passages from 2020 to 2021?

2) What are the similarities and differences in the results of the syntactic level text complexity of the NMET reading comprehension passages from 2020 to 2021?

3) What are the similarities and differences in the results of the discourse level text complexity of the NMET reading comprehension passages from 2020 to 2021?

3.2. Research Materials

Considering that this study makes use of the method of text analysis, it is crucial to discern and clean the data used in this research.

The texts involved in this research are all texts extracted from the NMET reading tests from 2020 to 2021, which include multiple choices question type (four options given) and matching type (five out of seven items).

The dataset contains NMET reading comprehension passages from 2020 and 2021, which are 76 in number. The final word count of Year 2020 is 13474 and that of 2021 is 10685.

Table2. *Corpus of NMET reading comprehension passages*

Year	# of exams	#of passages	#word count
2020	9 (I, II, III, Q-I, Q-II, BJ, JS, TJ, ZJ)	43 (4+4+5+5+5+5+5+5+5)	13474
2021	7 (Q-I, Q-II, Q-Jia, Q-Y, BJ, TJ, ZJ)	33 (4+4+5+5+5+5+5)	10685
Total	16	76	24159

Note: *I, II, III, Q-I, Q-II, Q-Jia, Q-Y is short for the national paper developed by NEEA. The difference is that, set I, II and III were used earlier than Q-I, Q-II, Q-J and Q-Y. Q-I, Q-II, were first used in 2020 and Q-J and Q-Y were first used in 2021. BJ, JS, TJ, ZJ is short for NMET reading comprehension passages developed by Beijing municipality, Jiangsu province, Tianjing municipality and Zhejiang province*.

3.3. Instruments

The instruments used in processing the NMET test texts are Coh-Metrix[1] and Eng-Editor[2], which are two online text analysis tools and provide indices in lexis, syntax and discourse. R studio[3] is an open source software for data analysis, as well as for producing charts and figures. Generally speaking, Coh-Metrix will provide the preliminary result of the 28 indices, such as word count and average word length. The other four indices: word beyond NMET syllabus, Eng-Editor lexical, syntactic and textual difficulty level will be extracted from Eng-Editor, and the last two indices, thematic contexts together with genres will be coded by the author manually according to the New English Curriculum Standards.

[1] cohmetrix.memphis.edu/cohmetrix2017
[2] https://www.languagedata.net/tester/
[3] https://www.rstudio.com

3.4. Research Procedure

Specifically, after collecting the 76 passages of 16 NMET reading tests, the author processed the texts first, such as deleted title and subtitle, as well as the Chinese characters, and encoded them into Word. Then the texts were put into Coh-Metrix and Eng-Editor to extract the indices at lexical level, syntactic level and discourse level (see Figure 1)

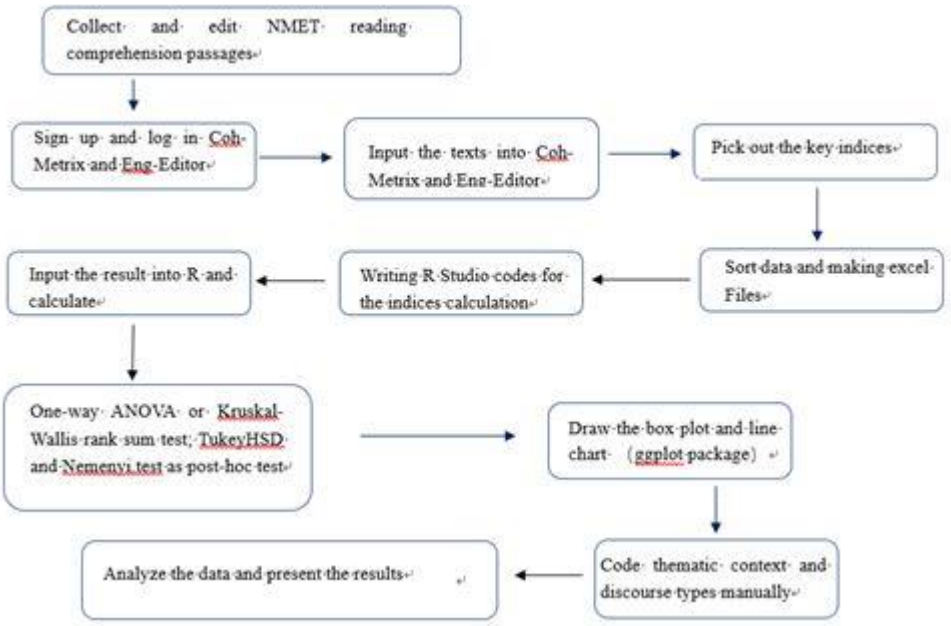


Figure1. *Flowchart of the research procedure*

The data were then put into excel files, because R Studio could not read file types such as word and txt. The values and groups were coded into "value" and "group", so the code could run. Since there were two years' data to be compared, the test methods utilized in processing this group of data are T.test and Wilcox.test, depending on whether the data was distributed normally. In addition to the quantitative statistic, this thesis also included two qualitative measures, which were thematic contexts and genres. The author manually classified these two indices thrice, with a two-week gap each time. She also enquired her peers for help classify the two indices, so the correctness of this part is proved to some extent.

4. RESULTS

The three research questions will be addressed and the implication will be discussed in this section.

4.1. Research Question One: Lexical Level Text Complexity

The result of the eight lexical complexity indices is presented in Table 3. As shown, there is no statistically significant difference in the five indices: word count, average word length, average word frequency for content words, type-token ratio, and word beyond NMET syllabus. Meanwhile, statistically significant difference exists among lexical diversity, i.e. MTLD (P=1.084e-13), VOCD (P=0.04), and also Eng-Editor lexical difficulty level (P=1.058e-13), which will be discussed next.

Table3. *The comparison result of lexical level text complexity indices of two years' NMET reading comprehension passages*

	2020		2021		*P*
	Mean	SD	Mean	SD	
L1	306.68	78.35	314.88	85.51	0.99
L2	4.26	1.03	4.58	0.36	0.78
L3	2.16	0.30	2.26	0.17	0.37
L4	0.56	0.06	05.5	0.07	0.79
L5	98.83	25.18	101.36	21.37	**1.084e-13**
L6	100.04	22.85	104.45	23.37	**0.04**
L7	4.95	2.56	4.64	2.94	0.45
L8	4.23	0.50	4.64	0.49	**1.058e-13**

Table 3 and the Figure 2 show that the mean MTLD of reading comprehension passage of NMET 2020 is 98.83 and that of 2021 is 101.36. Therefore, it could be concluded that the reading passages of NMET 2021 have a higher MTLD (P=0.04). Thus, the lexical diversity of reading comprehension passages of NMET 2021 is higher than that of NMET 2020.

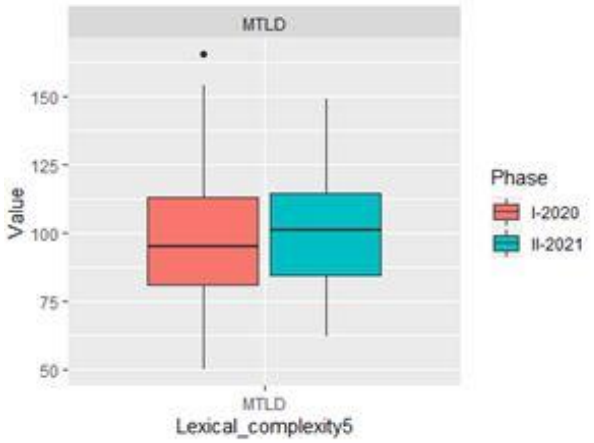


Figure2. *Boxplot of MTLD of two years*

Table 3 and the Figure 3 below present the result of VOCD of two years' NMET reading comprehension passages, which is another measure of lexical diversity. The mean VOCD of phase I is 100.04 and the mean VOCD of phase II is 104.45. Consistent with the result of MTLD, the VOCD of the two phases also appears to be statistically significant different (*P*=0.04). VOCD of reading comprehension passages of year 2021 NMET is also higher than that of 2020. Therefore, statistically, the lexical diversity of the NMET reading comprehension passages of 2021 is higher than that of 2020.

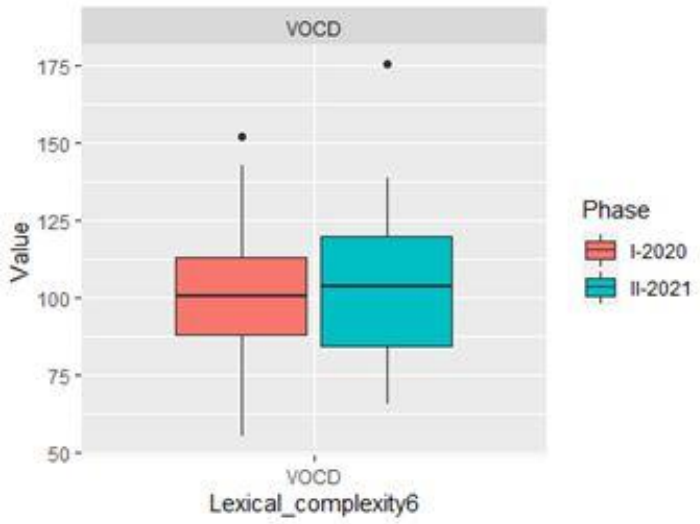


Figure3. *Boxplot of VOCD of two years*

In addition to MTLD and VOCD, statistically significant difference was also found in the index: Eng-Editor lexical difficulty level (P=1.058e-13). Overall, the mean value of Eng-Editor lexical difficulty level of year 2020's NMET reading comprehension passages is 4.23 and that of year 2021 is 4.64. The result may indicate that the Eng-Editor lexical difficulty level of 2021 is statistically higher than that of 2020. CSE has categorised nearly 3000 descriptors into 9 proficiency levels with 3 stages depicting the development of language ability (NEEA, 2018). Among which, candidates rated with level 1, 2 and 3 are at the elementary stage, while candidates of level 4, 5 and 6 are at intermediates stage and finally candidates at level 4, 5 and 6 belong to the advanced stage. So, according to the boxplot, for both two years, some NMET reading comprehension passages are over the level 6 or below 4, which

belongs to the difficulty level of the College English Test (CET) band 6[4] and the difficulty level of National Senior High School Entrance Examination (NSHSEE) respectively. This result might denote that the difficulty level of the NMET reading comprehension passages have a relatively large range.

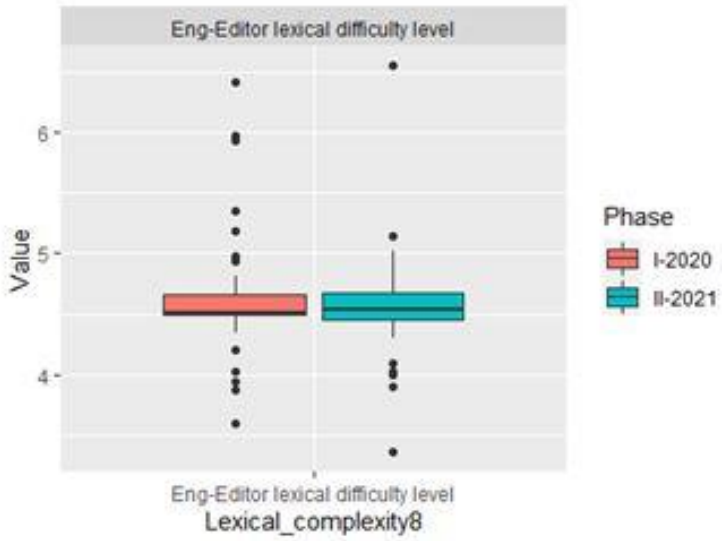


Figure4. *Boxplot of Eng-Editor lexical difficulty level of two years*

4.2. Research Question Two: Syntactic Level Text Complexity

After comparison, the result of the syntactic level text complexity of 2020 and 2021 is demonstrated in Table two. According to the data's distribution type, the author found that there is only one indicator, noun phrase density performs statistically significant difference, while the other ten all show no statistically strong distinction.

Table4. *The comparison result of syntactic level text complexity indices of two years' NMET reading comprehension passages*

	2020		2021		*P*
	Mean	SD	Mean	SD	
S1	16.06	3.52	15.50	3.47	0.64
S2	382.01	38.43	363.45	30.27	**0.02**
S3	1.29	0.18	0.88	0.17	0.46
S4	3.74	1.56	3.57	1.31	0.61
S5	5.82	4.82	6.19	6.12	0.84
S6	5.74	5.67	6.14	3.97	0.29
S7	0.47	0.15	0.41	0.17	0.13
S8	0.38	0.34	0.16	0.17	0.34
S9	0.17	0.07	0.16	0.06	0.25
S10	0.31	0.11	0.31	0.13	0.71
S11	4.05	0.67	4.32	0.60	0.008

Noun phrase density has shown statistically significant difference between year 2020 and 2021's NMET reading comprehension passages. With the mean noun phrase density of 382.01 and 363.45 respectively, 2020's noun phase density is significantly higher than that of 2021 according to the significant test method (P=0.02). The boxplot demonstrates that the noun phrase density of NMET reading comprehension passages developed in 2020 is higher than that of 2021.

[4] The College English Test, better known as CET, is a national English as a foreign language test in the People's Republic of China. It examines the English proficiency of undergraduate and postgraduate students in China. It includes two levels: CET4 and CET6 and enrolls millions of candidates each year.

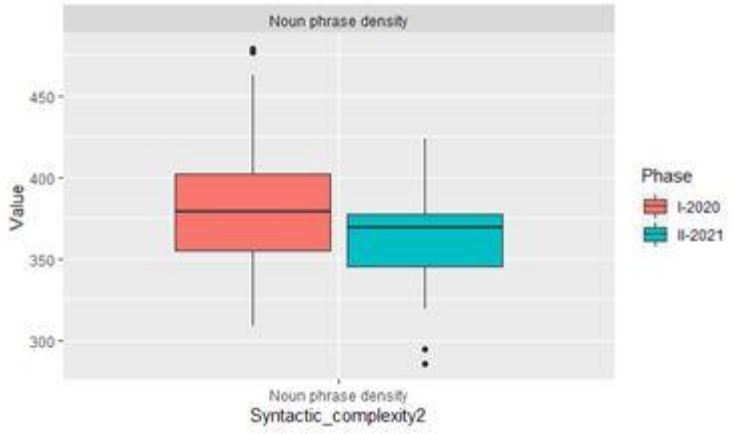


Figure5. *Boxplot of noun phrase density of two years*

4.3. Research Question Three: Discourse Level Text Complexity

Discourse level text complexity is composed of five parts: thematic context, genres, connectives, text abstractness and readability. After performing the significance testing method, the indicators show no statistically significant difference except average hypernymy for verbs (P=1.081e-13). In terms of thematic context and genres, because the dataset is not suitable for parameter test or non-parameter test, the author will present the descriptive analysis of these two sets of data.

Table5. *The comparison result of discourse level text complexity indices of two years' NMET reading comprehension passages*

	2020		2021		P
	Mean	SD	Mean	SD	
D-C1	23.62	9.96	24.54	9.16	0.68
D-C2	34.15	11.75	35.23	14.53	0.73
D-C3	13.93	8.35	15.51	9.06	0.45
D-C4	17.60	9.70	18.11	10.12	0.97
DA1	394.43	29.75	386.58	26.04	0.23
DA2	6.28	0.51	6.34	0.56	0.62
DA3	1.68	0.20	1.57	3.29	**1.081e-13**
DR1	63.47	9.86	63.10	12.72	0.92
DR2	8.71	1.96	8.25	2.35	0.70
DR3	15.51	5.99	14.70	4.35	0.50
DR4	4.59	0.41	4.57	0.44	0.48

Table6. *The thematic context of NMET reading comprehension passages of two years*

	Human and society		Human and themselves		Human and nature	
	2020	2021	2020	2021	2020	2021
BJ	2	2	2	2	1	1
JS	2	\	2	\	1	\
I	1	\	3	\	1	\
II	2	\	2	\	1	\
III	3	\	1	\	1	\
Q-I	3	2	2	2	0	1
Q-II	3	3	1	1	1	1
Q-J	\	2	\	1	\	2
Q-Y	\	4	\	0	\	1
TJ	2	2	2	2	0	0
ZJ	2	1	2	2	0	1
Sum	20	16	17	10	6	7
Total (76)	36(47%)		27(36%)		13(17%)	

Table 6 reports the thematic context involved in the reading texts of NMET 2017 to NMET 2021. Likewise, the proportion of the three categories of each phase is uneven. The overall trend is that in the two years that the proportion of human and society (47%) ranked first of thematic context. In contrast, human and themselves coupled with human and nature take account of 36% and 17% of the whole thematic context respectively. Nevertheless, the proportion of NMET 2007 in using the three categories of thematic context passages did not become better than that of 2020, with the proportion of human and society human and themselves still occupying most of the shares. All in all, the proportion of human and society is higher than the other two, especially human and nature, thus the distribution of the three categories being imbalanced.

Table7. *The genres of NMET reading comprehension passages of two years*

	Practical writing		Expository		Narration		Argumentation	
	2020	2021	2020	2021	2020	2021	2020	2021
BJ	1	1	1	2	2	1	1	1
JS	1	\	1	\	1	\	1	\
I	1	1	3	\	1	\	1	\
II	1	\	3	\	1	\	1	\
III	1	\	3	\	0	\	1	\
Q-I	1	1	3	2	1	2	0	0
Q-II	1	1	3	3	1	1	1	0
Q-J	\	1	\	3	\	1	\	0
Q-Y	\	1	\	3	\	0	\	1
TJ	0	1	0	1	0	1	0	1
ZJ	0	0	0	3	0	1	0	0
Sum	7	6	22	17	9	7	5	3
Total (76)	13(17%)		39(51%)		16(21%)		8(11%)	

The result in Table 7 indicates a common phenomenon of the reading comprehension passages the two years that the most frequently used genre is expository, taking up to 51%. The result might be attributed to the nature of NMET test, because it is such a large scale and high risk exam, and it has to be objective and avoid controversy, so expository is a “safe” choice. In contrast, argumentation tends to be least used in the five years’ NMET papers. While slight difference also exists among the genres involved in NMET reading comprehension passages of 2020 and 2021. For instance, the proportion of practical writing and narration is almost the same for two years (52%), but the proportion of argumentation of 2020 (11%) is higher than that of 2021 (9%).

Finally, the hypernymy for verbs is crucial in determining the abstractness of a passage. The hypernymy for verbs decreases from 1.68 to 1.57 (P = 1.081e-13) from Year 2020 to 2021, suggesting that the NMET reading comprehension passages of 2021 have become abstract verb concepts and therefore more difficult to understand from the perspective of text abstractness than the NMET reading comprehension passages of 2021. The boxplot below also shows the decreasing trend of this index from 2020 to 2021.

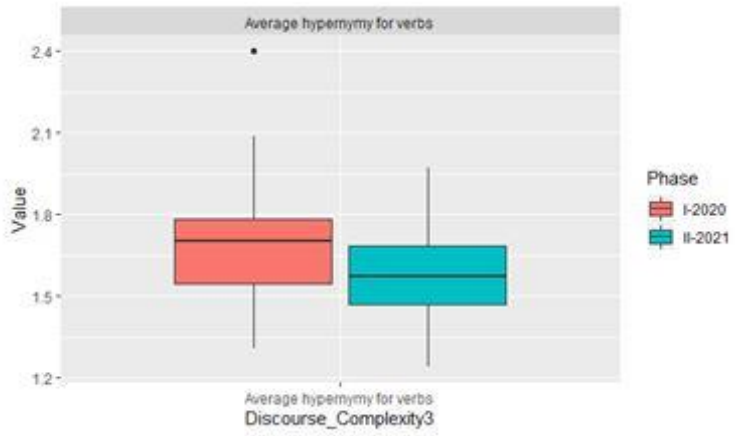


Figure6. *Boxplot of average hyperhymy for verbs of two years*

5. DISCUSSION

In the preceding section, the author compared the text complexity of the NMET reading comprehension passages between Year 2020 and 2021. The results of comparing the lexical complexity of the reading texts of the two phases suggest that there is no statistically significant difference in their word count, average word length, average word frequency for content words, type-token ratio and word beyond NMET syllabus. To rephrase it, the most direct measures of lexical complexity, word count, word length and word frequency imply no strong distinction after the reform. Nevertheless, statistically significant difference has been found among MTLD, VOCD, and Eng-Editor lexical difficulty level. MTLD and VOCD are the major predicators of lexical diversity, and both the two indicators in phase II are higher than phase I; therefore, it is proper to say the lexical diversity of phase II increased after the NMET reform.

In terms of syntactic complexity, only the indicator, noun phase density, evinces statistically significant difference in comparing the two years' data. The result might help to prove that the syntactic level text complexity of the two years' NMET reading comprehension passages is controlled reasonably. For example, the most direct indicator to measure syntactic level text complexity, average sentence length is found to fluctuate slightly over the past years, while overall the average sentence length of two years is 15 to 16. Only that the noun phrase density shows statistically significant difference in two years, the reason might ascribe to that in 2020, the NMET designers employed more passages with the topics of human and society, while using more expository to communicate this idea, while, these topics and genres naturally contains more noun phrases.

Thirdly, among the indices of discourse complexity, no statistically significant difference was found in the indices of connectives and readability. Nevertheless, the text abstractness index: average hypernymy for verbs shows statistically significant difference. Average hypernymy for verbs is an important indicator of text abstractness, and the decreasing trend of average hypernymy for verbs from 2020 to 2021 might hypothesize that the NMET reading comprehension passages in 2021 become more abstract. Combing the previous elaboration on noun phrase density, it could be summarized that there is a trend in the NMET designing process to choose and adopt the passages to make them more difficult to understand, and thus to assess the candidates' core literacy of English language.

Finally, the proportion of man and society of thematic context and expository of genres are still much higher than the other categories, the same as the proportion of the two in each year's reading texts. This imbalanced composition of these two categories should be improved in accordance of the *New English Curriculum Standards* and testing syllabus.

6. CONCLUSION

The major findings shed light on the significance of text complexity research of texts of NMET reading texts from multiple angles. The importance of reading comprehension in NMET would barely be weakened. Therefore, on the one side, the text complexity features, instruments and results spotted out in this study offer valuable empirical evidences for NMET designers and future study. One the other side, the high school teachers and students would also find the phenomenon this study found informative and therefore utilized in the daily learning and teaching.

Specifically, of NMET developing group, the selection and adaptation of texts are essential in the process of the whole test development. The NMET designers are recommended to take into account all three aspects of text complexity: lexical level, syntactic level and discourse level, and choose or adopt the NMET passages accordingly. For example, on the basis of this study, the author suggests when choosing the text for NMET reading it would be better to choose a text with a length of 225-300 words, a type-token ratio of 0.5, a MTLD and VOCD of 90-110 and a difficulty rating of Eng-Editor 4 to 5. Meanwhile, the syntactic complexity of the reading texts should increase the proportion of complex sentences and the low density of negative and passive voice sentences. The thematic context could have a more balanced frequency among the occurrences of human and themselves, human and society and human and nature, also with an appropriate increase in the proportion of argumentative essays. The NMET developing group could capitalize on the readability formula, for instance a text is more appropriate for NMET reading comprehension passages with the following parameters:

Flesh Readability at 50 to 60,

Flesh Kincaid Grade Level at 8 to 9,

Coh-Metrix L2 Readability at around 15 plus Eng-Editor text difficulty level at 4 to 5.

Meanwhile, the high school teachers of English are highly expected to carefully study the *New English Curriculum Standards*, the CSE and the NMET testing syllabus and other official documents, and make good use of the textbooks. At the same time, teachers could apply the traditional tools for calculating the readability of different passages, such as the Coh-Metrix and Eng-Editor, to determine text complexity of the texts used in teaching. Also, by that means, English teachers could choose appropriately graded reading texts for students, and cultivate students' reading proficiency and critical thinking. When teaching, the teachers could focus on examining the lexical diversity of the word the students mastered.

7. Limitations and Future Studies

There are two limitations due to realistic reasons. The data collected in thematic context and genres were coded manually. Although the author tried to conduct the coding process thrice, the data might be subjective to some extent. Secondly, the theoretical framework in this study is relatively inclusive; however, it has only measured a part of the features of text complexity, for instance, the factors of syntactic simplicity were not included in the framework. The main reasons contribute to that text features used in this study are too most closely related to text complexity and also because too many indices make it difficult to deal with all the texts.

Therefore, it is optimal for future studies to code the qualitative indices such as thematic context and genres to triangulate the results. In addition, future researchers could examine other indices to reveal a more integral picture of the text complexity of NMET, such as the syntactic simplicity and narrative it of passages. Conclusively, scholars are also encouraged to include diachronic datasets to enrich the research and verify the validity of NMET reform.

References

Chen, K., & Zhang, J. (2020). A study on ensuring the rating quality of the continuation task in the National Matriculation English Test. *China Examinations*, (12), 38-43.

Corkill, A. J., Bruning, R. H., & Glover, J. A. (1988). Advance organizers: Concrete versus 236 abstract. *The Journal of Educational Research*, *82*(2), 76-81

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115-135.

Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in A Foreign Language*, *26*(1), 92-113.

Far, M. (2016). The effects of text type, text length and text difficulty on vocabulary retention through grossing. *The Journal of Language Teaching and Learning*, *6*(1), 92-104

Green, A., Ünaldi, A.， & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy

of texts in tests of academic reading. *Language Testing*, *27*(3), 1-21.

Green, A., Ünaldi, A., & Weir, C. (2008). The cognitive processes of second language academic readers. LLAS Pedagogic Research Fund Project Report 11. Retrieved from: http://www.llas.ac.uk/projects.

Gu, X., & Wang, Q. (2008). Content validity study of the NMET reading comprehension tests. *Examinations Research*, (3), 102-114.

Gui, S. C. (1985). *Standardized testing - theory, principles and methods*. Guangdong Higher Education Press

Gui, S. C. (2007). *Gui Shichun's anthology on English language education*. Foreign Language Teaching Research Press.

Gui, S. C. (2017). *Selected academic papers of Gui Shichun.* Shanghai Foreign Language Education Press.

Guo, K., Jin, T., & Lu. X. (2018). Research and practice in text adaption: From readability formulas and multidimensional analyses to data-driven adaptions. *Foreign Language Testing and Teaching*, (3), 35-43.

Hu, X. (2018). A corpus-based quantitative analysis on linguistic features of reading comprehension section in College Entrance Examinations English Exam. *English Teachers*, (19), 40-42.

Huang, L., &Wang, J. (2020). Investigating reading text complexity of NMET based on python and Coh-Metrix. Foreign Language Testing and Teaching, (3), 1-11.

Jin, T., Lu, X., Guo, K., Li, B., Liu, F., Deng, Y., Wu, J., & Chen, G. (2018). Eng-Editor: An online English text evaluation and adaptation system. Guangzhou: LanguageData (languagedata.net/tester).

Khalifa, H., & Weir, C. (2009). Examining reading: Research and practice in assessing second language reading. Cambridge University Press.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris., & S. A. Stahl (Eds.), Children's reading comprehension and assessment (pp. 73-92). IEA Publishers.

Kontovourki, S. (2012). Reading leveled books in assessment-saturated classrooms: A close examination of unmarked processes. Reading Research Quarterly, 47(2), 153 - 171.

Li, W., Gui, S.C., & Li, X. (1989). Statistical procedures lation English Test. Foreign Language Teaching and Research, (2), 27-39.

Liu, Q. (2017). The development of the NMET over the past forty years. China Examinations, (2), 14-19.

Liu, Q. (2019). The strategies for protecting test fairness in China: Taking Gaokao as an example. *China Examinations*, (11), 1-6.

Lv, S. (2017). Reflections and theoretical analysis on the issues of forty years NMET in China. *Examinations Research*, *13*(6), 59-65.

Lyashevskaya, O., Panteleeva, I., & Vinogradova, O. (2021). Automated assessment of learner text complexity. *Assessing Writing*, *49*, 100529.

McCarthy, P.M., Jarvis, S. (2010). MTLD, VOCD-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42 (2), 381-392.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. Q. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

National Education Examinations Authority. (2018). China's standards of English language ability. Ministry of Education of People's Republic of China.

Perfetti, C. (2011). Decoding, vocabulary, and comprehension: The golden triangle of reading skill. In M. G. McKeown, L. Kucan, M. G. McKeown, L. Kucan (Eds.), *Bringing reading research to life* (pp. 291-303). Guilford Press.

Qi, X. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, *14*(1), 51-74.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* RAND Education

Tao, B. (2017). An exploration into item writing for the English achievement Test & NMET. *China Examinations*, (3), 25-44.

Wang, Q. (2018). Interpreting the six major changes of the Senior High School English Curriculum 2017. *Foreign Language Education in China*, (1), 11-19.

Xiao, Y. (2014). Research on content validity of reading comprehension in National Matriculation English Test Beijing (2010-2013). *English Teaching & Research Notes*, (1), 56-62.

Citation: *Yin Kailan. "Text Complexity of Reading Comprehension Passages in National Matriculation English Test: A Three-level Corpus Study" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 46-58. DOI: https://doi.org/10.20431/2349-0381.09S1006*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 59-67
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1007
www.arcjournals.org

A Preliminary Study on the Influence of Social Presence for Learning Satisfaction of Chinese Online Learners

Zhu Jingzhe

College of Foreign Studies, Jiaxing University, Zhejiang Province, China

Corresponding Author:** ***Zhu Jingzhe, *PhD of Mokpo National University, Korea, is director of Korean Language Department under College of Foreign Studies, Jiaxing University, Zhejiang Province, China. 13736403190@qq.com*

Abstract: *This study analyzes the effects of social presence on cognitive presence, emotional presence, and learning satisfaction, and attempts to empirically analyze whether learning readiness has a moderating effect on these relationships. Taking social presence as an independent variable and the relationship between emotional presence, cognitive presence, and learning satisfaction as the main research object, the author examined the effect of learning readiness as a moderating variable. The subjects in this study are Chinese undergraduates taking online course of Korean language on the designated learning platform. The questionnaire survey used Likert 5-level scale. A total of 189 valid questionnaires were collected, and SPSS was used to conduct exploratory factor analysis on the questionnaire structure. Smart PLS 3.0 was used to analyze the data structure. All the VIF values < 5. The conclusion thus obtained shows that social presence has both direct and indirect influence on learning satisfaction, and the overall effect is significant. Cognitive presence is a mediating effect on learning satisfaction, and emotional presence also needs to have a mediating effect through cognitive presence. Learning readiness moderates the relationship between social presence and emotional presence, and affects learning satisfaction through cognitive presence.*

Keywords: *social presence, cognitive presence, emotional presence, learning readiness, learning satisfaction*

1. INTRODUCTION

Scholars have different definitions of 'presence', but in general, the feeling of being in a place and belonging to a specific group is called presence (Anthony, 2002), which is different from online learning, where physical contact is excluded. In such a situation in a virtual space, learners have a qualitatively different understanding and experience than that in a traditional (offline) learning environment. Therefore, the learning experience, process and learning effect experienced by learners are different from offline learning in the traditional sense. As a variable to explain this issue, presence is brought back. Presence is referred to as the concept of the sense of being there, a kind of a subjective perception different from reality. When two people have a conversation in a space, the two people actually exist in the same space; however, the conversation situation or the presence of each other is very different. Even when learning the same online course, the respective levels of presence are different, the subjects are in physically different places, and in the learning environment in the ongoing virtual space, the question of presence naturally arises. According to Kim & Kang. (2010), this can be interpreted as the perceptually evoking fantasies or the second media experiences for representation.

2. THEORETICAL BASIS

The research regarding 'presence' called the Community of Inquiry Theory (CIT) was put forward by Garrison and Cleveland (2005). This theoretical model elucidates the behaviors and processes required for effective knowledge construction during online learning by describing three major presence, of which, Teaching Presence (TP) refers to the organization, design, facilitation of dialogue and direct teaching instruction; Social Presence (SP) refers to online discourse that promotes positive emotion, interaction and functional collaborative cohesion; Cognitive Presence (CP) refers to the degree to which learners construct meaning through conversation and reflection in an online learning community. The inquiry community model provides unique perspectives, methods and tools for online learning research, which has been widely recognized by researchers from all over the world (Jia & Li,

2020).

Other scholars have also described a community of inquiry as a TP, a CP, and a SP. While CP is referred to as the extent to which learners construct and verify meaning based on critical, continuous dialogue and reflection in an online inquiry learning community, SP dialogue and reflection in an online inquiry learning community, SP express their "true self" socially and emotionally through communication media in an online inquiry learning community. TP refers to the design, promotion and guidance of learners' CP and SP in order to realize the learning effect of learners' personal as well as educational value. The core of the inquiry community model is the superposition of TP, SP and CP. The superposition of the three presences is "deep and meaningful learning experience" (Lan，Zhong, Lv & Song, 2018).

Contrast to CIT, the Learning Presence Theory (LPT) holds that when a learner is faced with a specific learning situation, he or she feels an inner state through interaction with the environment. Wang and Kang (2006) divided learning realism into cognitive, social and emotional presence for discussion. This includes three overlapping and intersecting domains: cognitive domain, affective domain, and social domain. Teachers can use this model to describe each learner, and then design strategies for each individual. Students will not only have the opportunity to achieve their learning goals but will actively participate in the learning process. In doing so, they actually provide meaningful and engaging learning experiences for online learning students with diverse cultural and linguistic backgrounds (Wang & Kang, 2006)

Apart from this, LPT integrates factors from the affective domain with social and cognitive factors in the learning dynamics. In the social domain, the most important factor affecting learning and learning outcomes is social background. The learner's social background cultivates his personal characteristics, affecting the participation in group discussions and the group he belongs to. In fact, each learner has the specific background and culture of his own, and they inevitably reveal these characteristics during each learning endeavor. In this sense, consideration of social and cultural background is crucial (Wang, 2008)

In this study, presence was defined as learning presence. The social presence in learning presence was used as an independent variable. The main research subjects are college students who are learning online. They are the main learners of many online courses. The social scope of these learners is relatively fixed and the social background is relatively simple. Thus the social presence plays an important role in learners' course selection.

2.1. Social Presence (SP)

Defined as "perceived realism using communication media as a bridge", social presence (SP) is an important factor for non-face-to-face learning and can be felt through various non-face-to-face media such as computers, interactive TV, and mobile phones. According to Kim, Choe & Gwon (2014), which medium to use depends on the degree of perception of social presence. Social presence also refers to the degree to which learners have the sense of and express their psychological and social roles.

SP can also be interpreted as the learner's ability to identify themselves in their relationships with the community, to interact in a goal-directed manner in an atmosphere of trust, and to develop interpersonal relationships while appropriately expressing their individuality. The social presence perceived or expressed by a learner can determine psychological distance from other learning participants and influence the level of engagement in the learning process (Lee & Kim, 2015).

In virtual spaces, the desire to interact face-to-face is strong, which plays a decisive role in forming trust and affects whether more research or interaction will be conducted and the relationship between intimacy and immersion. Anonymity or virtuality created by non-face-to-face interactions can negatively impact overall attitudes; therefore, a method needs to be devised to minimize virtuality. Although direct face-to-face interaction cannot be provided in an online environment, it is necessary to provide a similar perception or feeling, the concept of a social presence that establishes elements of a learning environment for academic exploration and high-level interaction (Park, 2020). The researchers found that social presence creates a sense of intimacy through the exchange of opinions between learners and has a positive impact on learning outcomes. Furthermore, as online learning progresses, the amount of social information among the learners increases. This is positively

correlated with learning outcomes, suggesting that social presence can be a reason for improved learning outcomes (Kim, Son, Lee, Jeung, Jang & Lee, 2020).

Meanwhile, social presence is also understood as a sense of coexistence, influence, and cohesion, which provides a foundational process for learners to identify relationships with others and the communities that learners identify in the learning process. Especially in web-based learning, social presence increases intimacy between learners, which has a positive effect on academic satisfaction and plays an important role in achieving learning outcomes (Lee & Yun, 2012). Defining social presence as the degree of presence of an individual participating in the communication to the object with which it is communicated can also be broadly interpreted as the social relationship with the communication object, the degree to which people perceive others in the interaction, and find that each person is very sensitive to social presence. Here the perception of the senses is different, and such a difference plays an important role in the interaction (Kim & Cho, 2012).

2.2. Cognitive Presence (CP)

Cognitive presence is an element that reflects the knowledge experienced in the learning process. It is an essential element of authentic experience. The difficulty forming a high-level cognitive existence also supports such a notion that it is distinct from interaction. While it is true that cognitive presence is based on interaction, the simple exchange of information or the sharing of opinions does not necessarily produce presence. In addition, when interaction is structured and systematized as a communication characterized by reflective thinking and critical discourse, it may require a cognitive presence (Garrison & Cleveland, 2005).

2.3. Emotional Presence (EP)

Emotional presence refers to a kind of personal presence that allows users to freely express their emotions and feel comfortable through them via online or mobile communities. EP is defined as the degree to which learners become aware of themselves and generate positive emotions in their surroundings through contact with themselves in data and communication situations (Wang & Kang, 2006). In this sense, EP is also taken as acknowledging how much a person knows about their emotions as well as the freedom to express them in a learning situation. Within such a learning context, should EP be replaced with virtual reality, it would be interpreted as both the degree to which users can freely feel and express emotions and the degree of expression they feel when using digital media (Han, 2019).

2.4. Learning Satisfaction (LS)

Learning satisfaction, the most widely used primary measure of learning performance, refers to the learner's response to the satisfaction of the learners participating in the course with their learning. In the online learning environment, learners must learn independently. If they are satisfied with the education, the motivation to actively participate in learning will increase because the learning can achieve educational purposes. This is considered an important variable to measure learning performance (Joo, Ha, Yoo & Kim, 2010). LS is an important factor in acquiring knowledge as it allows learners to directly examine learners' responses to the classroom instructions and to know that learning satisfaction in an online learning environment is important for the improvement and development of effective teaching and other online education as well (Jeon & Yoo, 2020)

2.5. Learning Readiness (LR）

Learning readiness (e-Learning readiness) refers to the readiness of learners to be able to learn successfully in a learning environment because the learning opportunities provided by e-learning are the skills, cognitive strategies and motivations suitable for the new learning environment. This increases the chances of successful learning (Watkins, Leigh & Triner, 2004). Learning readiness includes the hardware readiness needed in terms of learning and of the status readiness for online communication and learning skills.

3. RESEARCH METHOD

The research method in this paper includes three parts: research model, research hypothesis, and the composition of the research objects and variables.

3.1. Research Model

The present researcher attempts to illustrate the learning process experienced by Chinese university learners in an online environment through presence, and to correlate learning outcomes with learning presence so as to verify the relationship between/among factors. In this study, of the three factors of learning presence recognized by Chinese university learners: SP, EP and CP, SP is the main factor. It is assumed that learners' satisfaction as a representative in terms of the relationship between learning outcomes.

The present research examines the appropriateness of the relationship between learning presence and learning outcomes in Chinese college students' online learning in the following:

To examine how SP, as an independent variable, affects other factors of LP, and ultimately affects learning outcomes;

To examine and compare how EP and CP, as mediators, affect desired learning outcomes, and

To examine whether the adjustment of learning readiness affects the relationship between presence and school effectiveness.

The research model designed shown in Diagram.1 below.

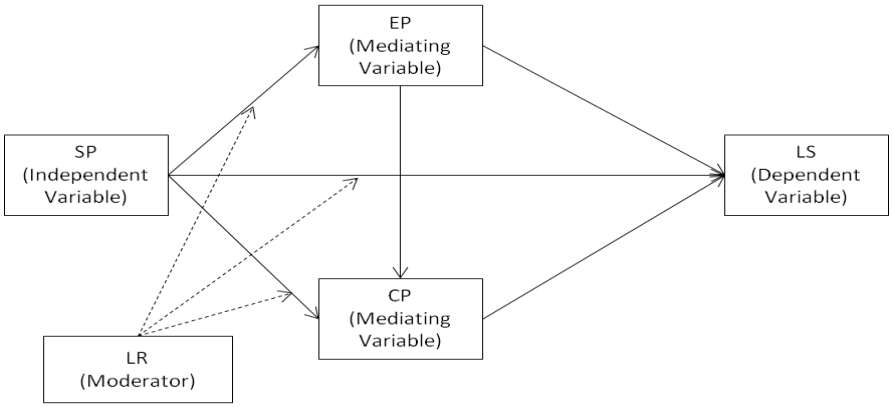


Diagram1. *Research Model*

3.2. Research Hypothesis

In the research model, social presence (SP) is used as an independent variable, emotional presence (EP) and cognitive presence (CP) are mediating variables, learning readiness (LR) is a moderating variable, and learning satisfaction (LS) is a dependent variable. Therefore, the research hypotheses are formulated as what follows:

Hypothesis 1: SP affects learning satisfaction.

Hypothesis 2: CP and EP act as a mediating role between SP and LS.

Hypothesis 3: LR will moderate the relationship between SP and LS.

3.3. Research Subjects and Variables

The subjects of this study are Chinese college students who take selected elementary Korean courses on the designated online learning platform, with no specific major and grade limit required. The prerequisite for answering the research questionnaire is that subjects have learnt offline no less than 8-hour preparation course. A total of 189 valid questionnaires were collected from these learners. Empirical studies (Wang & Kang, 2006; Kim & Kang, 2010) and the measurement tool of Shin and Chan (2004) were taken as reference.

The learning satisfaction in the present study is defined as the overall satisfaction of learners with online learning (see 2.4); Learning readiness refers to the measurement tool of Kim, Moon, and Par

K (2015). Based on the measurement tools validated in the pilot study, the author remodified and constructed the variables to suit the purpose of the present study, as diagrammed in Figure 2.

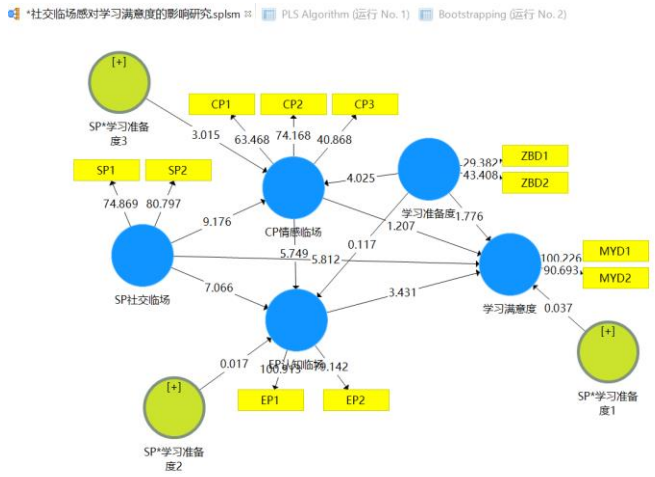


Figure2. *The model constructed by the author*

Where five factors were specified, namely social presence (SP), emotional presence (EP), cognitive presence (CP), learning readiness (LR) and learning satisfaction (LS). SP was used as an independent variable, EP and CP were used as mediating variables, LR was used as a moderator variable, and LS was used as a dependent variable (3.2.).

3.4. Questionnaire and Discussion

In order to ensure the reliability and feasibility of the study, the author designed a questionnaire using the 5-Likert scale, and conducted exploratory factor analysis on the 14 dimensions that the questionnaire might construct, as shown in Figure 3. The analysis of the data collected by the questionnaire is shown in Table 1 below.

Table1. *KMO and Bartlett Test*

KMOand Bartlett's Test		
Kaiser-Meyer-Olkin Measure Sampling Adequacy		.927
Bartlett's Test of Sphericity	Approx. Chi Square	1993.848
	df	91
	Sig	.000

From Table 1 above, we can see that the KMO value is very close to 1, indicating that the questionnaire we designed is very suitable for factor analysis. Another thing that draws our attention is the scree plot obtained from the questionnaire data, as shown in Figure 3, where goes the curve trend of the scree plot of this questionnaire.

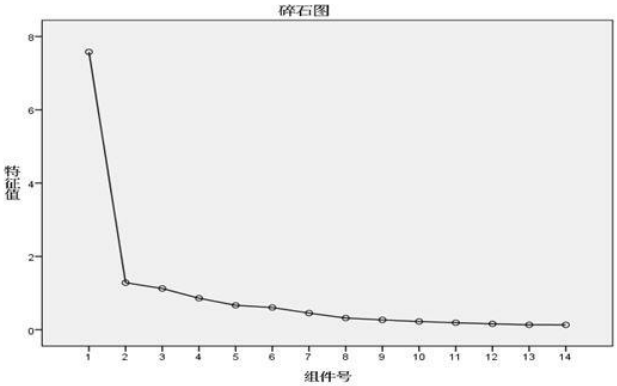


Figure3. *The scree plot of possible factors in the questionnaire*

We know that in exploratory factor analysis, the scree plot shows the possible number of dimensions constructed in the questionnaire designed. As shown in the figure above, of the 14 dimensions we constructed, starting from the seventh one on the abscissa, the curve tends to be gradually smoother downward. Our interpretation is that there may exist certain correlations between many variables in these dimensions we constructed, resulting in overlapping information, so that the latter dimensions become inconspicuous. Or we can also understand that more dimensions than necessary were set while the fewer factors under each dimension were specified.

4. ANALYSIS RESULTS

4.1. Reliability Test

In the reliability analysis of this study, both the Cronbach's Alpha value and the combined reliability exceeded 0.7. Therefore, the reliability can be confirmed (Fornell & Larcker 1981). Table 1 below shows the results.

Table1. *Reliability and Convergent Validity Test*

	Observed variable	Factor loading	Cronbach's Alpha	Combined r	Average Variance Extraction (AVE)
SP	SP1	0.94	0.865	0.937	0.881
	SP2	0.94			
CP	EP1	0.95	0.884	0.945	0.896
	EP2	0.94			
EP	CP1	0.91	0.904	0.94	0.838
	CP2	0.94			
	CP3	0.90			
LS	S 1	0.96	0.911	0.957	0.918
	S 2	0.96			

4.2. Validity Test

To verify the convergent and discriminant validity of the variables used in the study, the authors performed confirmatory factor analysis (CFA) using the PLS system[1]. The convergent validity of the PLS-based structural equation approach can be confirmed by factor loadings and standard errors (Hair, Ringle & Sarstedt, 2011). As shown in Table 1, the factor loading is above 0.7, the average variance extraction (AVE) value is above 0.5, and the combined reliability value is confirmed to be above 0.7. Meanwhile, in order to verify the discriminative validity, the AVE square root value and correlation coefficient value of each variable were compared, as shown in Table 2 below. The square root value of AVE is higher than the value of correlation coefficient between latent variables, and discriminant validity is judged to be valid (Fornell & Larcker, 1981).

Table2. *The Analysis of Discriminative Validity*

	EP	CP	SP	Satisfaction
EP	0.916*			
CP	0.827	0.947*		
SP	0.762	0.846	0.939*	
LS	0.776	0.843	0.863	0.958*

* AVE

4.3. Structural Equation Analysis Results

Structural analysis was carried out using the PLS algorithm. According to Hu and Bentler (1999), the Standardized Root Mean Square Residual (SRMR) value is an absolute measure that can be evaluated as a model fit criterion. The SRMR value was 0.054, which was lower than 0.08, and was judged to be

[1]PLS,i.e. Partial Least Square, is the latest achievement of partial least squares analysis and development of structural equation model analysis, forming a statistical analysis method of PLS-SEM. The latest version of Smart PLS software is Smart PLS3.0.

suitable. Apart from this, by examining whether there exists multi-collinearity among the analytical latent variables of the structural model, all the VIF values of LS between that of EP, CP, and SP turned out to be below the threshold value of 5 (Hair, Ringle & Sarstedt, 2011). This also confirmed that the explanatory variables in its basic assumptions are independent of each other.

4.4. Hypothesis and Testing

In this study, structural equation models were analyzed using Smart PLS 3.0 to test the hypotheses. Standard errors and t-values were calculated by the Bootstrapping algorithm using 5,000 subsamples (Kang & Hwang, 2021). Table 3 shows the resulting analytical values for the structural equation model path.

Table3. *Direct Effect Verification Results*

Paths	Path parameters	T value	P value	Results	R^2	f^2
SP ->LS	0.475	5.779	0.000	Sig	0.801	0.015
EP ->LS	0.106	1.288	0.198	InSig.		0.094
CP ->LS	0.301	3.487	0.000	Sig		0.292
SP ->CP	0.513	6.994	0.000	Sig	0.794	0.49
EP ->CP	0.433	5.699	0.000	Sig		0.312
SP ->EP	0.618	9.162	0.000	Sig	0.658	0.749

From the above table, we could see, as a direct effect of the structural equation, all are significant except the insignificant effect of emotional presence (EP) on learning satisfaction (LS). Social presence (SP) has a significant effect on LS (ß=.475, p=0.000), cognitive presence (CP) has a significant effect on LS (ß=.301, p=0.000).Our interpretation is that EP is very personal. Learners can express their emotions freely during learning. In the author's opinion, learners are aware of the emotional resonance around them through online learning situations. Such an emotional response has a herd effect, and their EP does not directly affect the judgment of the entire learning effect, but needs to judge and resonate with other learners' EP.

Table4. *Indirect and Total Effect Verification Results*

Paths	Path parameters	T value	P value	Results
SP ->EP->LS	0.066	1.218	0.223	Insig.
SP->EP->CP->LS	0.081	2.596	0.009	Sig
SP ->CP ->LS	0.154	3.405	0.001	Sig
SP ->LS（Total Effect）	0.776	16.771	0.000	Sig

As shown in Table 4, the overall effect of SP was significant through mediating variables. The indirect effect of SP on LS through CP is significant as well, but the effect of EP on LS turned out be not significant through the mediating variable, and the effect of EP on LS has, through the mediating effect of CP, significant impact on LS.

Table 3 shows that SP as an independent variable has a significant direct effect on EP and LS. The overall effect on satisfaction with learning shown in Table 4 was also significant, with a path coefficient of 0.776. SP is the degree to which learners feel and express their social roles, knowing themselves in their interactions with other learners, and develop interpersonal relationships while socializing with them. In this sense, SP is of great significance among Chinese learners, affecting cognition and emotion, and directly and indirectly affecting learning outcomes. When Chinese learners differ between their personal views and the general social cognition in online learning, their personal emotions follow the attitudes towards the learners around them and follow the public cognition, instead. Their personal emotions become blurred in the judgment of the final learning effect, thus unable to generate any direct impact.

Table5. *Moderating Variable Validation Results*

Paths	Path Parameters	T Value	P Value	Results
SP * LR ->LS	0.001	0.037	0.971	Insig
SP * LR ->EP	0.129	3.009	0.003	Sig
SP * LR ->CP	-0.001	0.016	0.987	Insig

The present research examines whether learning readiness (LR), as a moderator, moderates the relationship between SP and LS. Table 5 shows that LR is only significant for the moderating models of SP and EP. This research is an offline learning preparation process before formal online learning. The learning process includes the understanding of the network platform, the use of network tools, the preliminary understanding of Korean language and the primary pronunciation. The LR required by learners in online learning has a significant moderating effect between SP and EP, while the moderating effect between LS and CP is not significant. Therefore, it can be considered that the CP of Chinese online learners basically has a general understanding before and during the learning process, and the degree of readiness for learning cannot affect the above-mentioned cognition. A significant effect was formed under the adjustment of readiness.

5. CONCLUSION

This study analyzed the effects of SP on CP, EP, and LS. Furthermore, analyzed empirically was whether LR has a moderating effect on the above relationship. Through the empirical analysis, the following three conclusions are drawn:

At first, social presence (SP) has direct and indirect effects on learning satisfaction (LS), and the overall effect is significant. The importance of SP in online learning can be confirmed.

Next, cognitive presence (CP) is a mediating effect on LS, and EP also needs a mediating effect through CP. This means that even if the learner is either satisfied or dissatisfied with the learning process, the learning results has no substantial impact on the final LS. The final satisfaction hinges on CP.

Finally, learning readiness (LR) moderates the relationship between SP and EP, and affects LS through CP. In order to increase LS, cognitive and emotional factors need to be considered at the same time. This shows that it is imperative to strengthen training and understanding in teaching design, software and hardware operations, learning methods, etc. What is more, improving the degree of preparation before formal learning is conducive to the smooth learning process, thus to upgrade the level of learner satisfaction.

6. LIMITATIONS

Two limitations remain. The first one is that the sample size is not big enough, and the analysis results did not take into account the subjects' gender, major and grade's perception of the learning content, situational awareness, and the learner's experience in the learning situation. And no detailed analysis was presented of whether any impact on the level of perception, emotion and social sensitivity. The second limitation is that, although 14 dimensions was constructed in the questionnaire, exploratory factor analysis showed that the actual construction of the latter 7 dimensions was not obvious.

REFERENCES

Anthony, G. (2002). Beyond student perceptions Issues of interaction, presence and performance in an online course. JANL 6(1), 21-40.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Researc, 18(1), 39–50.

Garrison, D. R., & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough.American Journal of Distance Education, 19 (3), 133-148.

Hair, J. F., Ringle, C. M. & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet, Journal of Marketing Theory Practice, 19(2), 139-152.

Han, Kwang-Seok.(2019).The Effect of Presence and Interactivity of Digital Signage Using 3D Virtual Reality on Brand Experience and Attitude, Journal of Digital Convergence,17(4), 299-307.

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

Jeon, Seong, J. & Yoo, Hyo Hyun. (2020). Relationship between General Characteristics, Learning Flow, Self-Directedness and Learner Satisfaction of Medical Students in Online Learning Environment. The Journal of the Korea Contents Association, 20(8), 65-74.

Jia,L.F.&. Li, H.L. (2020). Influence of presence on learning cognition of online learners: Analysis of conditional process based on community of inquiry theory. e-Education Research. 41(2), 45-51.

Joo, Y. J., Ha, Y. J., Yoo, J.W., & Kim, E.K.(2010).The structural relationship among teaching presence, cognitive presence, social presence, and learning outcome in Cyber University, Journal of the Korean Association of Information Education, 14 (2), 175 – 187.

Kang, M.J. & Hwang, D. H. (2021). A Study on the effect of SNS characteristics on festival loyalty: Focused on the mediating effect of festival flow and the moderating effect festival involvement. Academic Society of Event & Convention, 17 (1), 21-44.

Kim, G., Choe H. & Gwon, S. (2014).Influence of social presence on online community users' continuance intention. The Journal of the Korea Contents Association, 14 (2), 131 – 145.

Kim, J. & Cho, H. (2012). An influence of a sense of classroom community and social presence on learning satisfaction in a Cyber learning setting, Journal of the Korea Academia-Industrial Cooperation Society, (13) 8.

Kim, J. M. & Kang, M. (2010). Structural relationship among teaching presence, learning presence, and effectiveness of e-learning in the corporate setting. Asian Journal of Education, 11(2), 29-56.

Kim, J. M., Son, K. T., Lee, P. E., Jeung, J. Y., Jang, H. B. & Lee, H. J. (2020). The Effects of Interaction between instructor-student and student-student on learning achievement in synchronous e-learning for major classes for university students: The mediating role of learning flow. Journal of Agricultural Education and Human Resource Development, 52(3), 30-33.

Kim, S.R., Moon, E., & Park, I. (2015). Investigation on the relationships among students' e-learning readiness, teaching presence and learning effects in an online learning environment. Korean Association for Educational Information and Media. Korean Association for Educational Information and Media doi:10.15833/kafeiam.21.4.687.

Lan, G. S., Z hong, Q.J., Lv, C.J., & Song, Y.T. (2018). Exploring relationships between learning presence and community of inquiry model. Open Education Research, 24 (5) 92-107.

Lee, E. & Kim,Y.J.(2015). Differences in the social presence of instructor by instructor's social intervention and its effects on learning satisfaction in an online university, The Journal of Korean Association of Computer Education, 18(3), 69 – 78.

Lee, J. & Yun, N. (2012).The effects of task value, Twitter self-efficacy, and Social presence on learning satisfaction on Twitter discussion, Journal of the Korean Association of Information Education, 16 (1), 51 - 60.

Park, Gyeongwon. (2021). The effects of teaching reality and learning reality perceived by college students on learning satisfaction in non-face-to-face classes, The Journal of the Korea Contents Association, 21(12), 175 - 181.

Shin, N. M., & Chan, J. (2004). Direct and indirect effects of online learning on distance education. British Journal of Educational Technology, 35(3), 275-288.

Wang, M.J. (2008). A Cybergogy Model for Engaged Learning, Open Education Research， 14 (2), 22-27. 2008.

Wang, Minjuan & Kang, Myunghee.(2006).Cybergogy for engaged learning: A framework for creating learner engagement through Information and communication technology, 225-226.

Watkins, R., Leigh, D., & Triner, D. (2004). Assessing readiness for e-learning. Performance Improvement Quarter, 17(4), 66-79.

Citation: *Zhu Jingzhe. "A Preliminary Study on the Influence of Social Presence for Learning Satisfaction of Chinese Online Learners" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no.s1, 2022, pp. 59-67. DOI: https://doi.org/10.20431/2349-0381.09S1007*

International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 9, Special Issue 1, May 2022, PP 68-77
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
https://doi.org/10.20431/2349-0381.09S1008
www.arcjournals.org

Cultural Differentiation and Job Performance: the Moderation Role of Feedback Seeking Behavior

Liu Chang

Alliance Manchester Business School, University of Manchester, Britain

Corresponding Author:** ***Liu Chang, *Alliance Manchester Business School, University of Manchester, Manchester M139PL, Britain. changliu306@ hotmail.com*

Abstract: *This study concentrates on an investigation on how the variables of feedback seeking behavior (FBS) and job performance in the cross-cultural work settings. The author predicts that the feedback seeking from the sources of supervisors, colleagues and organization will influence on the relationship between the cultural different characters of collectivism and individualism. The empirical study conducted in the multinational companies located in China supported the assumptions. This research is innovative in theory and methodology which enriches managerial literature by exploring more inclusive solution that benefit work outcomes in cultural diversity teams using effective moderator to interact with the typical, influential cultural practice characters representing the salient characters of diversity typologies of separation.*

Keywords: *culture separation job performance work feedback seeking moderation*

1. INTRODUCTION

Cultural diversity in work setting has become an increasing point of discussion and concern associated with job effectiveness (van Knippenberg & Schippers, 2007). The researchers assert the discrepancy in findings regarding diversity-related outcomes (van Knippenberg & Schippers).

Some studies have indicated a number of benefits related to a heterogeneous workplace associated with positive organizational outcomes (Cox & Blake, 1991; Kearny et al., 2009). Nevertheless, there are also numerous researchers indicating that the phenomenon detrimentally associates with the negative effects of a demographically diverse workplace. Addition to these "analogous to a double edged sword" (Millikens & Martins, 1996), a meta-analyses review concludes that the team member diversity is far from exerting and a statistically significant effect on work performance (Joshi & Roh, 2007).

In order to respond to these challenges of the complexity, mixture manifestations of diversity findings in reviews of research, many researchers have promoted elaborating possible contingency factors to clarify the connection between within-unit diversity and the unit-level outcome (Van Knippenberg, et al., 2007; Kearney, et al.,2009).

Furthermore, factors moderating the relationship between work group diverse and work outcomes, eg, diversity beliefs (Meyer & Schermuly, 2012), national variety (Ayub & Jehn, 2014); organizational identities (Few & Joshi, 2013); shared objectives (van Knippenberg et al., 2011); diversity climates (Lauring & Selmer, 2011); psychological safety (Singh et al., 2013), and etc. have the potential to substantially contribute to the effective management of workforce diversity.

Derived from Van Knippenberg and Schippers (2007), the team cultural diversity is roughly conceptualized in this paper as the distribution of differences among members in cultural attribute, resulting in a unique mixture, which affects an individual's behavior, attitude, assumption, and expectations. Cox (2001) underlines that diversity is reflective of the variation in social and cultural identities among people existing together in an employment setting. Moreover, according to Gorman (2000), diversity may be conceived of as the varied perspectives and approaches members of different identity groups bring to the workplace.

Hofstede (1980) defines national culture as the collective programming of the mind which distinguishes the members of one human group from another‖. In management, culture is defined in terms of values, beliefs, norms, attitudes and behavioral tendencies that are used to develop cultural categories (Javidan, et al., 2006). Recognized as an important organizational variable, culture remains the construct that is difficult to catch, and the complex, and diverse elements. Bailey, et al., (1997) suggested that "culture affects individual desire for, behavior toward and perception of performance feedback". The author of this paper does not propose a main effect hypothesis between cultural differentiation and job performance in the present research, rather, the author aims at exploring the potential links between critical diversity approaches and diversity management interventions in organizations focusing on a pair of specific cultural context variables that the author believes are associated with cognition patterns and work outcomes. The chosen variables were taken from the well-known Global Leadership and Organizational Behavior Effectiveness (GLOBE) study (House, et al., 2004).Using a theory-driven approach, the author identified two of well-known cultural variables: collectivism and individualism that would explain behaviors of individuals, groups and organizations across different countries (Bond, 1996; Earley & Erez, 1997; Hofstede, 2001; Triandis, 1995), as such would have clear links to the relationship between cultural differentiation and job performance.

The author has a close examination on how some factors of feedback seeking behavior variable moderate cultural separation and job performance in the cross-cultural work setting. The author predicts that the feedback seeking from the sources of supervisors, colleagues and organization is associated with the relationship between the culturally different characters of collectivism/ individualism and job performance. The author contributes to the management literature by investigating on how feedback seeking (FBS) behavior, a noticeable variable in organizational behavior research, specifically, the channels of feedback seeking, links to the relationship between the culturally practical separation characters of collectivism and individualism and job performance.

2. THEORETICAL CONCEPTION AND HYPOTHESIS DEVELOPMENT

Cultural Differentiation, defined as a composition of differences incultural attribute among unit members. Built on social categorisation (Turner, 1987), cultural differentiation suggests unfavourable effects of team consequence since diverse team tends to divide itself into distinguished subgroups via the social categorisation processes generating relationship conflicts and impeding collaboration, which in turn increase turnover among team members. On the other hand, 'contact hypothesis' (Allport, 1954) can benefit in enhancement of problem solving, decision making (cf. Homan, et al., 2007) and innovation and creativity in management. The author considers the value dimension of individualism and collectivism as cultural differentiation that are different in cultural category.

Collectivism is defined as - the degree to which organizational and societal institutional practices encourage and reward collective distribution of resources and collective action. (Javidan et. al, 2006)".

Individualism corresponds to the primacy people place on themselves over their aggregate social group (Mary &Steven, 2000). The conceptualization of individualism-collectivism (IC) has been shown as two independent factors with two contrasting poles opposing points on a continuum (Earley& Stubblebine, 1989; Hofstede, 1980). Most conceptualizations and measures of IC reflect its multidimensional and multilevel construct (Earley & Gibson, 1998; Earley et al., 1999; Schwartz, 1990; Triandis, 1995).

FBS behavior refers to the initiative of individuals to seek valuable information about their job performance in order to better adapt to the organizational environment and individual development needs (Ashford & Tsui, 1991). FBS behavior is considered as important and valuable self-fulfillment and interpersonal insightfulness that are relative with emotional intelligence to employee performance and managerial practice (Goleman, 1998, Ashford, et. al, 2003). The employees who actively seek feedback not only have a high degree of identification with the organization, but can quickly integrate into the organization and show good performance at work (Ashford, et al, 2007).The performance of FBS behavior is affected by its own characteristics. More studies show that the higher the frequency of individual feedback seeking, the better the communication and the higher the job satisfaction and performance, and vice versa (Tepper et al., 2006). Feedback giving and feedback seeking are integral and reciprocal activities (London, 1997). Furthermore, feedback giving occurs across levels of analysis. Individuals engage in feedback-seeking source choices and they can seek the information

from any channel available. (Ashford & Tsui, 1991; Callister et al., 1999). In this study, the author identifies three FBS sources: from the supervisor, colleagues and from organization. (e.g., unit reports, posted charts). Because feedback giving is affected by cultural characteristics in the cross cultural setting, the solicitation of feedback might be impacted as well.

To date, though the cultural diversity, feedback, and job performance have been heavily examined factors in the organizational behavior research respectively, the reviews show that researchers either examined the relationship between the culture and FBS behavior, for example, the theme of discussion is primarily that the strategies of FBS behavior are influenced by the individual's cultural (Mary & Steven, 2000) and how culture might affect organizational feedback giving, the recognized antecedent to an individual's feedback-seeking behavior (Ashford & Cummings, 1983; Mary & Steven, 2000), or much of this literature is concerned with showing the effectiveness of feedback on influencing future individual behavior and performance (Cusella, 1987). This study focuses on how the relationship between the ways values are expressed as reflected in the characteristics of cultural practice and job performance is directly influenced by the interpersonal interaction and communication that occur in each culture. Correctly using FBS behavior can bridge the gap between the cultural diversity team and wok outcomes.

Since collectivist employees are loyal and obey their leader, they are favored by the leaders and maintain better relationship, However, identified such concepts as face saving (Earley & Erez, 1997; Triandis, 1990) collectivist individuals would defend their egos by engaging in the defensive strategies avoiding openly asking their supervisors, they tend to seek feedback from their leaders privately.

With information about themselves, the individuals in individualistic cultures, in which direct communication is valued, would be more likely to express clear, direct require for feedback (Aycan & Kanungo, 2001). Therefore, it should be natural for the employees that are considered high in individualist orientation. To directly ask for feedback with leaders to address their personal information needs.

Reviews have showed that the superiors pay more attention to those who have active FBS activities, but neglect those who lack FBS action. Openly seeking feedback upward would be useful to help the leader to know more about the member so that the leaders would pay close attention to the member and give the members the valid information which is conducive to the members' task fulfillment. Therefore, it is based on these assumptions that the author of this paper proposes the following hypotheses:

H1a. FBS from supervisor will be positively associated with performance for the individuals that are considered high in collectivist orientation.

H1b: FBS from supervisor will be positively associated individuals that are considered high in individualist orientation.

Image defense or preventing embarrassment is important for employees who are more in collectivist orientation (Kim & Nam, 1998; Triandis, 1990). Moreover, in high collectivist societies, direct criticism is avoided because team harmony is essential (Fletcher & Perry, 2001). The employees who are high in individualistic cultures tend to be more concerned with preserving their own ego, whereas people in collectivist value is more lenient with others' party. That is, in collectivist cultures, face giving (i.e., allowing room for the other person to maintain or recover his or her face) is important (Ting-Toomey, 1999). Face-saving and face-giving behaviors focus less on the accuracy of a statement and more on what is culturally appropriate for the context (Samovar, et al., 2006). Lenience biases are likely to limit the willingness of cultures with more of a collective identity people to provide direct negative feedback to others.

In an individualistic culture, employees would prefer finding out needed information for themselves directly upward as mentioned above rather than by inquiring from peers owing to their competitive relationship. Hence, the author proposes:

H2a. FBS from colleagues will be negative association with performance for the individuals that are

considered highly collectivist-oriented.

H2b: FBS from colleagues will be negative association with performance for the individuals that are considered highly individualism-oriented.

For employees who are in high collectivist societies attach great importance to the interests of the organization and abide by the organizational rules and norms. So the intention to know the individual record in the organization is normal, the emphasis will be on the betterment of the organization through improving accuracy and understanding the task (Trope, 1982). The collectivists exhibit more organization-monitoring FBS behavior than individualists (Hwang and Francesco, 2010). In collective culture, the feedback process is focused on the formal structure to create a method of self-enhancement, (Shackleton & Ali, 1990), and the organizational level feedback is the routine like the direct reports', statistical records about the workforce of the unit. And with evaluating the position the individual is in the achievement in the organization, the individual can perform better, enhance their advantage and avoid their shortcoming. This can increase work quality.

In individualism culture, the feedback process is less focused on formal structure (Shackleton & Ali, 1990) and there is an intrinsic belief in individual decisions (Kluckhohn & Strodtbeck, 1961), and, thus, individual goals become the primary focus of behavior (Triandis, 1990).

Hence the author postulates:

H3a: FBS from organization is positively related to the culture considered high in collectivist orientation.

H3b: FBS from organization source is less positively associated with performance in individualistic orientation.

3. METHODOLOGY

Data were collected from 427 participants from multinational corporations located in China including joint ventures, corporation sole. To achieve sufficient statistical power for the multilevel modeling, the current data for analysis were chosen from countries that had at least 10 participants following the criterion based on simulation studies (Kreft & DeLeeuw, 1998). These participants were from 8 countries having sample sizes that ranged from 10 to 81employees. Totally, 553participants including supervisors and subordinates respectively came from eight countries, and 427 questionnaires were collected, accounting for 76%response rate owing to well organized. Based on these 427 individuals, a preliminary analysis revealed that75.3% were male. The average age was 31.65, with the ages ranging from 22 to 58. To test the hypothesis, the author obtained the records of the employees' performance evaluations from each organizations.

Feedback Seeking Sources

The author used the questionnaire developed by Vande Walle et al. (2000) with questions as how frequently participants sought feedback from supervisors, colleagues and organization source regarding (a) overall job performance, (b) technical aspects of the job, (c) values and attitudes of the corporations, (d) role expectations, and (e) social behaviors using the 5-Liket scale anchored by 1 (almost never) and 5 (very frequently). The Cronbach's alpha for this scale was .92.

Cultural practices. The Globe data were taken (House et al. 2004 for individualism; Javidan, et al., 2006 for society collectivism). The scores of collectivism and individualism were used separately in this study though the two conception were usually considered as in opposite pole. Because the author thought that it could not divided the score of individualism and collectivism for an individual by half to half.

Table 1 lists the 8 countries included in this study and the scores for each of the countries collectivism and individualism as well as the means and standard deviations for FBS sources of supervisor, colleague organization.

Analytic Strategy

Using Mplus 4.0 (Muthén & Muthén, 2006), multilevel modeling was run to test hypotheses (see, table 3) since each member in the sample was nested under the corresponding country culture from which he or she came, avoiding underestimate or overestimate standard errors for parameter estimates. (Bryk & Raudenbush, 1992; Snijders & Bosker, 1999). To test the main effects of FBS sources on cultural variables, the author estimated an intercept-only regression model for FBS sources at Level 1 and predictive effects of cultural practice variables on the Level-1 random intercept were Level-1 random regression slopes of collectivism and individualism in predicting cultural value both by group means when they were entered into corresponding multilevel models to avoid interpretation difficulties and spurious findings according to Hofmann and Gavin's (1998) suggestion. In addition, to provide a realistic view of how these cultural practices operate in concert with each, the author included them together simultaneously as Level-2 predictors in the analysis.

4. RESULTS

Means, standard deviations, and correlations among all variables across countries are presented in Table 1.

Table1. FBS Seeking from Three Sources and Cultural Practice Scores Included in the studies

	Feedback Seeking Sources							
	Colleagues		Supervisor		Organization		Collectivism	Individualism
Country	*mean*	*sd*	*mean*	*sd*	*mean*	*sd*	M	M
1. China	3.73	0.33	3.85	0.36	3.89	0.39	4.77	3.23
2. France	3.75	0.41	3.89	0.35	3.89	0.32	3.93	4.07
3. Italy	3.83	0.42	3.91	0.35	3.87	0.32	3.68	4.32
4. Japan	3.85	0.40	3.68	0.29	4.06	0.33	5.19	3.13
5. South Korea	3.70	0.36	3.67	0.42	4.02	0.36	4.40	2.80
6. Singapore	3.75	0.37	3.69	0.37	4.23	0.35	4.90	3.10
7. Sweden	3.93	0.39	3.80	0.35	4.10	0.38	4.10	2.78
8. USA	3.87	0.36	3.93	0.41	3.85	0.37	4.20	3.80

Table 2. Correlations, Reliability, Means, and Standard Deviations for the Study Variables

			Correlation					
Variables	M	SD	1	2	3	4	5	6
1.Supervisor source	3.85	0.45	—					
2.Colleagues Source	3.81	0.41	.07	—				
3. Organization Source	3.83	0.37	.09*	.6	—			
4. Collectivism	3.72	0.36	.017**	.05	.32**	—		
5. Individualism	3.61	4.45	.28**	.07	. 。。 .	.08	.03	—.
6.. Job Performance	5.61	1.16	.36*	.05	.31**	.04	.02	—

Note: N = 427 for the correlations between individual-level variables (Variables 1-3) and country-level variables (Variables 4–5). To calculate them, the author assigned the same country-level scores to all individuals who were from the same country. The correlations between two country-level variables were calculated at the country level with a sample size of N=8 *p < .05. **p < .01.

Testing Cross-Level Effects of Moderation Effects of Feedback seeking sources Variables

To test the cross-level moderation hypotheses, the author estimated two multilevel models that examined how FBS sources from supervisors, colleagues and organization predict the relationships between cultural context variables and job performance. Specifically, the author entered collectivism and individualism value respectively, as the Level-1 predictor of job performance in two models. The author then entered three feedback sources variables as the Level-2 predictors in predicting the

random intercept (0) and random slope (1) from the Level-1 regression. The coefficients of feedback sources variables (i.e., γ11, γ12 and γ13) provided the test of our hypotheses. As indicated in Table 3, collectivism was significantly related to the job performance slope, γ11 = 0.082, z = 2.76, p< .01.

Table 3. Multilevel Models for Testing Cross-Level Moderation Effects of FBS Sources

	Level-1 prediction			
	Collectivism → JP		Individualism → JP	
Model	Coefficient	SE	Coefficient	SE
	Random intercept model without Level-2 predictors			
Level 1—variance(σ^2_{within})	0.106**	0.006	0.106**	0.006
Level 2				
random intercept (β_0)				
Intercept (γ_{00})	3.675**	0.226	3.67** 5	0.226
Variance (τ_0)	0.018**	0.005	0.018**	0.005
Random slope (β_0)				
Intercept (γ_{10})	0.079 *	0.035	0.096*	0.051
Variance (τ_1)	0.009*	0.005	0.010**	0.004
	Random intercept and slope model with Level-2 predictors			
Level 1—variance (within2)	0.106**	0.006	0.106**	0.006
Level 2				
Random intercept (β_0)				
Intercept (γ_{00})	3.648**	0.232	3.650**	0.232
Supervisor source (γ_{01})	0.106*	0.057	-0.057	0.057
Peers source(γ_{02})	-0.057	0.048	0.106*	0.048
Organization source (γ_{03})	0.018	0.032	0.018	0.032
Residual variance (σ^2_{e0})	0.012**	0.005	0.012**	0.005
Random slope (β_1)				
Intercept (γ_{10})	0.679	1.465	1.301	0.924
Supervisor source (γ_{11})	0.082**	0.032	0.185 *	0.087
Colleagues source (γ_{12})	0.047	0.186	0.005	0.086
Organization (γ_{13})	0.117**	0.057	0.126**	0.058
Residual variance (σ^2_{e0})	0.005	0.017	0.003	0.005
Variance in random intercept accounted for by Level-2 predictors (%)	65.0		68.8	

Note. Level-1 *N*=427 ; Level-2 *N*=8 *p < .05. **p < .01.
Predictors at Level 1 were centered by group means.

The author plotted the significant interaction at conditional variable of feedback source of supervisor (i.e. ±SD) following Cohen, et al., (2003) procedure. As shown in Figure.1, Feedback seeking source from supervisor was positively associated with the relationship between collectivism and performance, supporting H1a. A similar finding was obtained that feedback seeking from organizational source– job was positively and significantly related to the collectivism – job performance slope, γ13=0.117, z= 2.16, p< .05, performance slope, γ13=0.117, z= 2.16, p< .05, colleagues did not show positive effects in predicting collectivism –job performance slope, providing support for H2a.

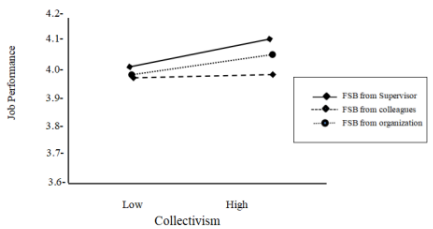


Figure 2.1.1. Influence of feedback seeking sources on collectivism- job performance

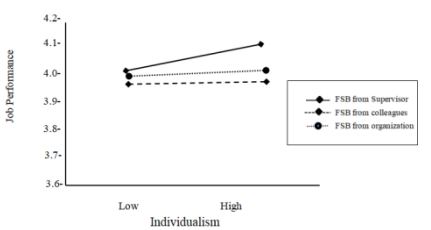


Figure 2.1.2. Influence of feedback seeking sources on individualism- job performance

As shown in Figure 2, the relationship between individualism value and job performance is stronger

when feedback seeking from supervisor, supporting H1b. The author also found that feedback seeking from either colleagues or organization was negatively related to the relationship between the individualism value and job performance supporting H2b and H3b.

Then, comparing the residual variances of the current models with nested models that did not contain the cross-level interaction term of feedback seeking sources variables (i.e., the random intercept and slope model without Level-2 predictors in Table 3), the author found that as a set of predictors, feedback seeking sources explained 67.0% of the variance in the random collectivism–job performance slope and 58.8% of the variance in the random individualism–job performance slope. Overall, these findings suggest that feedback sources are related to the relationships between cultural variables and job performance.

5. CONCLUSION

The author used a theory-driven approach to assess the way in which feedback seeking sources variables might affect the relationship between cultural context variables and job performance. The finding that the relationship between collectivism and job performance was higher for the members who frequently seek feedback from their supervisors and organizations. It is noticeable that the relationships between source from supervisor and job performance were decreased in countries characterized by high collectivism after organization source variable is controlled for. This makes sense because collectivism cultures are characterized by people abide by organizational norms, prioritizing organizational goals over individual goals. On the other hand, the results of that also show that members who have individualist value directly expressing their true thoughts and feelings regardless of their status or power relationships. In such cultures, members are more likely to know how their leaders evaluate their work, their strengths and weaknesses, and make clear the goals of their future efforts. Consistent with the hypotheses, the author found a main effect for colleagues feedback source had not been found on both individualism and collectivism suggesting a leniency bias in collectivism cultures and competition in individualist cultures.

Implications: Researchers have been trying to find ways to address the negative effects of cultural differences on job performance and turn them into positive factors using intervention methods such as moderating variables. This study helps us to have a deeper understanding of the potential links between critical cultural diversity approaches and diversity management interventions in organizations associated with cognition patterns and work outcomes.

It seems that the relationship between a cultural diverse workforce and its job performance is much more complicated. This is primarily due to whether cultural differentiations mean ”conflict” or “contact” depending on of factors and processes moderating the link between diversity and its potential benefit and costs (Qin et al., 2012). That the extent to which workforce heterogeneity will have a beneficial or detrimental effect on group performance depends on how the heterogeneous groups are managed within an organization (Kochan et al.2003). Cultural differentiation in diverse work force can be managed well for an organization associated with many benefits if a number of factors like feedback seeking sources variables are taken into consideration and addressed in an effective way. Feedback seeking behavior is considered as an important element for workforce individual to improve their job performance. However, among factors of feedback seeking factors (eg. Srrateg), the FSB sources are the critical elements needed to be addressed. Since asking leaders’ opinions about oneself’s performance can gain more useful feedback and correspondingly receive guidance even supporting which is benefit for job performance. As shown in this research, individuals in workforce cultural heterogeneity seek feedback due to the different cognitive backgrounds, mental-models, experiences, and perspectives brought by team members from different cultural backgrounds (Cox & Blake, 1991; Kearny et al., 2009).

Tough FBS from supervisor will be positively associated with performance for both the collectivists and individualists. The individualists seem to be easier to access to the information from leaders that is advantage to job performance due to their bold, directly requiring. While the collectivists who actively seek feedback from the organization will benefit in improving job performance effectively. Comparing the leader source and organization source, seeking feedback from colleagues only is the supplement only.

The author contributes to the managerial literature on investigating on how feedback seeking behavior

(FBS) which is noticeable variable in cultural diversity organizational behavior research, specifically, the channels of feedback seeking, links to the relationship between the cultural characters of collectivism and individualism and job performance.

REFERENCES

Allport, G.W. (1954). The mature of prejudice. Addison-Wesley.

Ashford, S. J., Blatt, R. & Vande Walle, D. (2003).Reflections on the looking glass: A review of research on feedback-seeking behavior in Organizational Journal of Management, 6:773-799.

Ashford, S. J., George, E., & Blatt, R. (2007). Old assumptions, new work: The opportunities and challenges of research on nonstandard employment. Academy of Management Annals, 1(1), 65–117. doi: 10.1080/078559807.

Ashford, S. J., & Tsui, A. S. (1991). Self-regulation for managerial effectiveness: The role of active feedback seeking. Academy of Management Journal, 34: 251–280.

Ashiord, S, J. & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. Organizational Behavior and Human Performance, 32: 370-398.

Aycan, Z., & Kanungo, R. N. (2001). Cross-cultural industrial and organizational psychology: A critical appraisal of the field and future directions. In N. Anderson, D. Ones, H. K. Sinangil, & C. Viswesvaran. (Eds.). Handbook of industrial, work, and organizational psychology: Volume1. Personnel psychology. Thousand Oaks, CA: Sage.

Ayub, N., & Jehn, K. (2014). When diversity helps performance: Effects of diversity on conflict and performance in workgroups. International Journal of Conflict Management, 25(2), 189–212.

Bailey, J., Chen, C., & Dou, S. (1997). Conceptions of self and performance-related feedback in the U. S., Japan, and China. Journal of International Business Studies, 28, 605– 625.

Bond, M, H, & Smith, P, B. (1996). Cross-cultural social and organizational psychology. Annual review of Psychology, 47: 205-235.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Sage.

Callister, R.R., Kramer, M. &Turban, D.B. (1999). Feedback seeking following career transitions. Academy of Management Journal, 42: 429-438.

Cogliser, C. C., Schriesheim, C. A., Scandura, T. A., & Gardner, W. L. (2009). Balance in leader and follower perceptions of leader member exchange: Relationships with performance and work attitudes. Leadership Quarterly, 20: 452-465.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). Applied multiple Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). Applied multiple for the behavioral sciences (3rd ed.). Hillsdale, NJ: Erlbaum.

Cox, T. H. (2001). Creating the multicultural organization. A strategy for capturing the power of diversity. Jossey Bas.

Cox, T. H., & Blake, S. (1991). Managing cultural diversity: Implications for organizational competitiveness. The Academy of management executive, 5(3), 45–56.

Cusella, L. P. (1987). Feedback motivation and performance. In F, M., Jablin, L. L., Putnam, K. H., Robert & L, W. Porter. (Eds.). Handbook of organizational communication: 624-768. Sage.

Earley, P. C & Erez, M. (1997). New perspectives on international industrial/organization psychology. New Lexington Press.

Earley, P. C, & Gibson, C. B. (1998). Taking stock in our progress on individualism-collectivism: 100 years of solidarity and community. Journal of management, 24: 265-304.

Earley, P. C, & Stubblebine, P. (1989). Intercultural assessment of performance feedback. Group and Organizational Studies, 14: 161-181.

Few, W. T., & Joshi, M. (2013). Top management team decision making: The role of functional and organisational identities on the outcomes of TMT diversity. European Journal of International Management, 7(1), 56–73.

Fletcher, C., & Perry, E. L. (2001). Performance appraisal and feedback: A consideration of national culture and a review of contemporary research. Handbook of Industrial, Work and Organizational Psychology, 1. 126-144.

Goffman, E. (1959). The presentation of self in everyday life. Goffman, E. (1959). The presentation of self in everyday life.

Goleman, Daniel. (1998).What Makes a Leader. Harvard Business Review

Gorman, F. (2000). Multinational logistics: Managing diversity. Air force journal of logistics, 14(3), 10–43.

Harrison, D.A. & Klein, K. J (2007) What's the difference? Diversity construct as separations or disparity organizations. The Academy of Management Review. Vol. 32, No. 4 (Oct., 2007), pp. 1199-1228

Hofmann, D., & Gavin, M. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. Journal of Management, 24, 623– 641.

Hofstede, G. (1980). Culture's consequences：International differences in work-related values. Newbury Park CA: Sage.

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. (2nd Ed.). Thousand Oaks, CA: Sage

House, R., Hanges, P., Javidan, M., Dorfman, P., & Gupta, V. (2004). Culture, leadership, and organizations. Thousand Oaks, CA: Sage.

Hwang, A. & Francesco, A.M. (2010).The Influence of individualism — collectivism and power distance on use of feedback channels and consequences for learning. Academy of Management Learning & Education, 9 (2): 243-257.

Javidan, Mansour, Peter, W., Dorfman, Mary Sully de Luque & Robert, J. House. (2006). In the eye of the beholder: Cross cultural lessons in leadership from project GLOBE source. Academy of Management Perspectives, Vol. 20, No. 1 (Feb., 2006), pp. 67-90.

Joshi, A., & Roh, H. (2007). Context matters: A multilevel framework for work team diversity research. In J. Martocchio. (Ed.). Research in personnel and human resource management (Vol. 26, pp.1–48). JAI Press.

Kearny, E., Gebert, D., & Voelpel, S. C. (2009). When and how diversity benefits teams: The importance of team members' need for cognition. Academy of Management Journal, 13(3), 581–598.

Kim, J. Y., & Nam, S. H. (1998). The concept and dynamics of face: Implications for organizational behavior in Asia. Organization Science, 9, 522–534.

Kluckhohn, F., & Strodtbeck, F. (1961). Variations in value orientations. Evanston, IL: Row, Peterson.

Kreft, I. I., & de Leeuw, J. (1998). Introducing multilevel modeling. Sage German History 14(2): 244-246.

Lauring, J., & Selmer, J. (2011). Multicultural organizations: Does a positive diversity climate promote performance? European Management Review, 8(2), 81–93.

London, M. (1997). Job feedback: Giving, seeking, and using feedback for performance improvement. Lawrence Erlbaum Associates.

Marry, F. Sully De Luque, & Steven, M.Sommer. (2000).The impact of culture on feedback-seeking behavior: An integrated model and propositions. Academy of Management Review. Vol. 25. No. 4. 829-849.

Meyer, B., & Schermuly, C. C. (2012).When beliefs are not enough: Examining the interaction of diversity fault lines, task motivation, and diversity beliefs on team performance. European Journal of Work and Organizational Psychology, 21(3), 456–487.

Milliken, F. J., & Martins, L. L. (1996). Searching for common threads: Understanding the multiple effects of diversity in organizational groups. The Academy of Management Review, 21(2), 402–433.

Muthén, L., & Muthén, B. (2006). Mplus (version 4.0) [Computer software]. Los Angeles: Author.

Paik, Y. (2016). Multilevel conceptualization of leader–member exchange processes: A comprehensive review. Social Behavior and Personality: An International Journal, 44, 401-413.

Samovar, L. A., Porter, R. E., & McDaniel, E. R. (2006). Intercultural communication: A reader. Belmont, CA: Wadsworth.

Schwartz, S. H. (1990). Individualism-collectivism: Critique and proposed refinements. Journal of Cross-Cultural Psychology, 21: 139-157.

Shackleton, Viv. J. & Ali, Abas. H. (1990). Work-Related Values of Managers. A Test of the Hofstede Model. Journal of Cross-Cultural Psychology 21(1):109-118.

Singh, B., Winkel, D. E., & Selvarajan, T. T. (2013). Managing diversity at work: Does psychological safety hold the key to racial differences in employee performance? Journal of Occupational and Organizational Psychology, 86(2), 242–263.

Snijders, T. A., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage

Tepper, B. J., Duffy, M.K., Henle, C.A., & Lambert, L.S. (2006). Procedural injustice, victim precipitation, and abusive supervision, Personnel Psychology, 59, 101-123.

Ting-Toomey, S. (1999). Communicating across cultures. Guilford.

Triandis, H. C. (1990). Cross-cultural studies of individualism and collectivism. In J. J. Berman (Ed.). Nebraska symposium on motivation, vol. 37: 41-133. University of Nebraska Press.

Triandis, H. C. (1995). Individualism and collectivism. Boulder, CO: West viewer.

Trope, Y. (1982). Self-assessment and task performance. Journal of Experimental Social Psychology, 18: 201-215.

Turner, J. C. (1987). Toward a cognitive redefinition of the social group. In Henri, Tajfel (ed.). Social identity and intergroup relations (p 15-40). Cambridge University Press.

VandeWalle, D., Ganesan, S., Challagalla, G. N., & Brown, S. P. (2000). An integrated model of feedback-seeking behavior: Disposition, context, and cognition. Journal of Applied Psychology, 85, 996–1003.

Van Knippenberg, D. & Schippers, M.C. (2007). Work group diversity. Annual Review of Psychology, Vol. 58, pp. 515-541.

Van Knippenberg, D., Dawson, J. F., West, M. A., & Homan, A. C. (2011). Diversity fault lines, shared objectives, and top management team performance. Human Relations, 64(3), 307–336.

Citation: *Liu Chang. "Cultural Differentiation and Job Performance: the Moderation Role of Feedback Seeking Behavior" International Journal of Humanities Social Sciences and Education (IJHSSE), vol 9, no. s1, 2022, pp. 68-77. DOI: https://doi.org/10.20431/2349-0381.09S1008*

ISSN 2349-0373

ARC Publications

37-1-4(15), First Floor, Second Line,
Annavarappadu, Ongole, Andhra Pradesh, INDIA, PIN-523001.
info@arcjournals.org