JAEPR, *4*(1),
Copyright © 2019,
ISSN: 1930-9325

RESEARCH ARTICLE

Premises and Challenges of High-Stakes Examinations: National Higher Education Entrance Examination Mathematics Test Scores in China

Chunlian Jiang
University of Macau

Do-Hong Kim
Augusta University

Chuang Wang
University of North Carolina at Charlotte

Jincai Wang
Soochow University

International comparative studies suggest that students from countries with high-stakes examinations often perform better than students from other countries. Nowadays more and more countries, including the United States, are implementing high-stakes examinations at the national or state levels. In this paper, we use the National Higher Education Entrance Examination (NHEEE) in China as an example to illustrate the premises and challenges of high-stakes examinations. NHEEE, commonly known as *Gaokao*, is the only measure used in China to determine if and which college a high school graduate is admitted to. This study examines the reliability and validity of scores obtained from the 2014 mathematics test of this critical examination that determines the future of thousands of students in China. Results of the Rasch analysis indicated that the unidimensionality assumption was tenable. The results also showed that the item reliability and separation were satisfactory, but the person reliability and separation were low. The low person separation reliability indicates that the exam is not sensitive enough to distinguish between low- and high-performing students. Examination of the person-item map suggested a need for more items at the intermediate and difficult levels to improve the reliability and validity of the test scores and to match the students' ability levels. Results showed that the majority of items displayed little or no DIF between male and female students. Predictive aspects of validity are also reported.

Keywords: high-stakes examination, national higher education entrance examination, Rasch analysis, reliability test, validity test.

International comparative studies like the Trends in International Mathematics and Science Study (TIMSS) indicate that students from countries with high-stakes examinations often perform better than students from other countries (Woessmann, 2001). Using data from TIMSS 1995, Woessmann (2001) found that students in countries with centralized examinations scored 16 points higher in mathematics. Since then, more and more countries, including the United States and Australia, implemented national assessment of students' performance even though there may be more cons than pros for high-stakes examinations (Madaus, 1991).

High-stakes examinations have been criticized for their inappropriate use of results (Wu & Hornsby, 2012), for their undesirable "backwash" or "trickle-down" on classwork and study of students at lower grades, and for their negative effect on students' personality characteristic (Madaus, 1991). However, it is still widely used in many countries including China, India, Singapore, South Korea, Japan, etc. National Higher Education Entrance Examination (NHEEE), commonly known as *Gaokao*, is important in China for high school graduates because the results could determine not only what university they can enter, but also their future career (Davey, Lian, & Higgins, 2007; Lambert, 2015; Wang, 2006). NHEEE is the only measure used by university admission offices in China to evaluate their applicants. The scores of NHEEE determine whether a student can be admitted to a prestigious university or not and whether a student can study in the major he/she chooses. The job opportunities for college graduates are not the same for all college majors or institutions. Quite a large number of high school graduates could not have the opportunity to pursue higher education because of their low scores in NHEEE. In 2016, about 9,400,000 students took NHEEE on June 7-9 (China News, 2016), while only about 6-7% of them (approximately 600,000 students) could be admitted to the so-called prestigious universities (Xiong, 2016). It is not exaggerating to claim that NHEEE scores play a large determining factor in a student's career choice, so families in China invest a lot of time and money into their children's preparation of NHEEE.

The pressure even moves downward to elementary schools. Many families stop their fifth-grade children's after-school activities, such as music and sports, so that their children can have time to attend after-school and/or weekend courses in the major academic areas, such as Chinese, mathematics, and English, hoping that their children could get admitted to prestigious middle schools. To these families, prestigious middle schools prepare their children for top-tier high schools, in which the students are more likely to be admitted to prestigious universities (Yang, 2006). As students move up each grade level, pressure increases because they are getting closer to taking NHEEE. Despite the fact that NHEEE plays a central role in students' education and future career success in China and that the interpretation and fair use of test scores are the most important component of validity (Messick, 1995), few studies have investigated the validity of these scores. The present study aims to fill this gap.

THEORETICAL FRAMEWORK AND RELATED LITERATURE

National Higher Education Entrance Examination (NHEEE)

NHEEE can be dated back to the *Keju* Examination in Tang dynasty (618-907) in China as a national civil service examination (Feng, 1995; Yu & Suen, 2005; Zhang, 1988). Since then until the Qing dynasty (1616-1911), Chinese government uses *Keju* to select talented officials to serve the country. Nowadays, NHEEE is not only used for college admission for high school graduates, but it is also used to judge the job performance of teachers and administrators and the overall quality of schools (Lambert, 2015).

NHEEE is used as the only measurement for Chinese students' abilities to get admitted not only to local universities but also to universities abroad. For example, NHEEE is suggested to serve as a gold standard to accept Chinese high school graduates for undergraduate programs in universities in Australia (Olsen, 2009). Up to 1,000 universities in 14 nations accept NHEEE test scores as an admission criterion (Zhang, 2015). Though the NHEEE cut-offs are different across different provinces, the cut-offs for the top tier universities in individual provinces account for the top 10% of the candidates, which is about 5% of the age population. For this reason, Olsen (2009) suggested to use NHEEE cut-offs to process the applications from Chinese high school graduates for undergraduate programs in Australia. Universities in the United States are also testing the practice to enroll Chinese high school graduates into colleges based on their NHEEE results (Schultz, 2015). Though the admission rate in China has increased dramatically in the past two decades from 27.3% in 1990 to 87% in 2012 (China Education Yearbook Editorial Board, 2014), many parents want their children to pursue higher education in the top-tier universities, which creates great tensions among parents and their children (Davey, Lian, & Higgins, 2007; Liu & Wu, 2006).

NHEEE is also important in terms of the ranking of high schools and university tiers their graduates can get admitted to. Almost all schools take serious actions to have a high proportion of their students entering higher education institutions (*Shengxue Lv 升学率*), which is generally taken as a measure of the quality of a high school. The actions include: (a) Accelerating the teaching pace to finish it earlier in order to free some time for review, which may affect the consolidation of what they are learning; (b) Reducing courses (e.g., music, art, and physical education) and/or topics that are not included in NHEEE, which may prevent the development of students' interests and talents in arts as well as their physical wellness; (c) Administrating examinations/tests frequently, which may make students tired of the examinations and hurt their psychological well-being; and (d) Selecting good teachers to teach seniors, which may weaken the teaching efficiencies of students at lower grades (Yang, 2007). This undue emphasis on *Shengxue Lv* happened as early as in 1963 (Yang, 2007). The great pressure on the students and their teachers have been criticized not only by educators but also by the public, media, and universities in China as well as in other countries (Cockain, 2011; Liu & Wu, 2006; Yang, 2007).

Since the independence of the People's Republic of China, NHEEE has been revised several times. The institution that is responsible for the design of NHEEE changed from individual universities in 1949 to the university unions in the northwest, northern, and Eastern parts of China in 1950-1951. Since 1952, the Ministry of Education took over this responsibility and power. Many regulation rules for NHEEE were set up in 1952-1957. For example, starting from 1954, the Ministry of Education in China releases the examination syllabus for NHEEE every year to specify the objectives and nature of NHEEE as well as the content coverage and requirements for students to systematically review the topics and to prepare for the examination (Yang, 2007). NHEEE was interrupted by the Cultural Revolution (1966-1976) and resumed in 1977. Since the resumption, NHEEE developed very smoothly. In 1985, Shanghai became the first region that was allowed to develop its own examinations for high school graduates. In 2002, Beijing, Guangdong, and Henan were also permitted to adopt independent examinations. Starting from 2004, authorities in more provinces were allowed to have their own examinations in line with the general guidelines including examination syllabi set by the Ministry of Education. Based on the national examination syllabi, the examination authorities in individual provinces release supplementary documents to explain the requirements in more details, including the proportions of problems in different formats, problems at various difficulty levels, and even sample tests. In 2014, for example, more than 20 sets of mathematics examinations of NHEEE were used in China. Questions have been raised about the quality of these examinations that were locally developed and administered in individual provinces. As a result, the national unified examination (NHEEE) became increasingly used across provinces again in recent years. However, there seems limited research that examines the validity of scores obtained from the NHEEE. Up to the present, there are very few technical reports on the reliability, validity, and even basic psychometric properties of the NHEEE (Hu, Li, & Gan, 2014; Wang, 2006; Wu, 2007). Limited evidence on the validity of NHEEE scores presents a challenge for educators, policymakers, and researchers as they undertake major reform efforts, based upon the NHEEE scores, in the examination and admission system in China. The current study aims to fill this gap by evaluating the psychometric properties of the NHEEE scores in mathematics, one of the three core subjects that all students have to take.

NHEEE Mathematics Examination

The NHEEE mathematics test is intended to test "two basics" (Chu, Wang, Wang, & Ding, 2005). Two basics refer to basic mathematical knowledge and basic mathematical skills, which are fundamental characteristics of Chinese mathematics education (Zhang, Li, & Tang, 2004). Basic mathematical knowledge includes topics that are important in secondary mathematics, for example, function, equation, inequalities, conics, vectors, and trigonometric functions (Chu et al., 2005). Basic mathematical skills that are intended to be tested in the NHEEE mathematics test include logical thinking skills, computational skills, spatial imaginary skills, data handling skills, and creative and application skills (Ding, 2008). The NHEEE mathematics test is also intended to test students' understanding of basic mathematical ideas and fundamental mathematical methods (Chu et al., 2005). Basic mathematical ideas include functional and equational ideas, integral ideas of numbers and graphs, classifying and combination ideas, transformation ideas, ideas about special cases and their generalizations, finite vs. infinite ideas, ideas about certainty and uncertainty. Fundamental mathematical methods include cutting and patching methods, proof by contradiction methods, methods to determine coefficients by substituting givens into several equations, and methods of substitution (Chu et al., 2005). Items are designed to examine students' basic knowledge and skills and their understanding of basic mathematical ideas and fundamental methods; however, it is very hard to say that an item will test only one or two of them. Even for a very simple item like "Given that $A=\{x\,|\,x^2-2x-3\geq 0\}$, $B=\{x\,|\,-2\leq x<2\}$, *find* $A\cap B$", the problem solver has to understand the meaning of the symbols, be able to solve the quadratic inequality, and finally be able to draw a number line to find the intersection of the two sets. Logical thinking, computational, and spatial imaginary skills are needed. Therefore, we shall first look at the content areas the test items covered and then explain their relative difficulties from the thinking skills tested.

Validity Evidences of NHEEE

Despite the fact that mathematics is one of the core subjects in NHEEE, the extant research on the NHEEE mathematics exam is limited both in number and in scope. Researchers have obtained very conflicting results with regard to its prediction level for students' further academic performance (Hu, Li, & Gan, 2014; Wu, 2007). Hu et al. (2014) found that students' performance in the three core subjects and the comprehensive test of NHEEE 2005 were all significantly correlated with their college academic performance. However, Wu (2007) found that the correlation coefficients between mathematics scores in NHEEE 2002 and academic achievement in the four-year study in universities were low, and even negative in half of the eight universities in science and engineering. Wu also found that the correlation coefficients between mathematics scores in NHEEE 2002 and academic achievement for students in mathematics majors were very low. These inconsistencies across studies and the contradictory findings make it difficult to draw valid conclusions and make generalizations. In the United States, however, numerous studies have been conducted about the validity of the Scholastic Aptitude Test (SAT), which is used as one of the requirements to U.S. universities (e.g., Beard & Marini, 2015; Patterson & Mattern, 2011, 2012, 2013a, 2013b).

In the popular book entitled *China's Yearbook 2014 of National Entrance Examinations to College: Mathematics*, Zang and Sun (2014) classified the item difficulties of the 24 items into low, intermediate, and high levels. Nine items (items 1-4, 7, 13-15, and 17) were identified at low level, ten items (items 5-6, 8-10, 18-19, and 22-24) were at intermediate level, and the rest (items 11-12, 16, and 20-21) were at the high level. They classified the items based on their experiences in teaching of high school mathematics and students' learning but not on any empirical data. Additional research is warranted to investigate the technical properties of the NHEEE mathematics (Wang, 2008) as researchers in the United States did for the SAT. This study will provide useful information for universities to make valid inferences concerning student abilities and to make important decisions for students, such as admission.

The theoretical framework that guided this study is Messick's (1995) framework of validity. Validity is the degree to which a test measures what is supposed to measure, so the interpretation and use of the scores (consequential validity) is of utmost importance. The classical models of test validity include content validity, criterion-related validity, construct validity, consequential validity, and that criterion-related validity has two forms: concurrent validity and predictive validity (Gay, Mills, & Airasian, 2009). Messick challenged these classical models of test validity and viewed validity as a single unitary construct. He argued to include consequential validity for the meaning and interpretation of the test scores and the fair use of the test scores are the most important component of validity, this is particularly important for the use of scores students obtained in NHEEE, which were now used as a unique measure for the recruitment of higher education institutions.

Gender Differences in Mathematics Performance

Mathematics is often taken as a male domain (Hyde, Fennema, Ryan, Frost, & Hopp, 1990). As the gender gap in mathematics performance narrowed in recent years (Hyde & Mertz, 2009; OECD, 2013), Andreescu, Gallian, Kane, and Mertz (2008) argued that female students with high ability in mathematics can be identified and nurtured. However, comparative studies across countries suggest that male students outperform female students in more countries (OECD, 2016).

Similar trends can be found in China. In 1957, only about 23% of the students in higher education institutions were female. From 1995 to 2004, the proportion of female students increased from 35.4% to 45.7%. The year of 2007 is the first year with more female students than male students admitted to higher education institutions (*Wuhan Wanbao*, 2012). This trend kept going in recent years with the difference in proportion of female and male students approaching 10% in

2013. It is interesting to note among the 63 students who received the highest scores in NHEEE in 27 provinces in 2012, 33 (52.4%) were female (*Wuhan Wanbao*, 2012). The proportion of female students whose scores were the highest in the respective provinces increased from 29% to 53% from 1952-1999 to 2000-2015 (Airuishen China's University Alumni Association, 2016). More female students than male students were studying in the Department of Mathematics in both top-tier universities and normal universities in China (Yang, He, & Ning, 2010). In a provincial normal university, the ratio of male and female students is 1:3.67 (Song & Zhang, 2017). Female university students also performed better than male students in mathematics (Qiu, Chen, & Xiao, 2009). So how did male and female students compare in high school mathematics? The results from previous studies were inconsistent. Tian and Zhu (2014) used the NHEEE 2011-2012 data from Ningxia province and found that male students performed better than female students in mathematics. However, Ye (2011) used the NHEEE 2006-2010 data from Zhejiang province and found that female students performed better than male students in mathematics in 2006-2008, but worse than male students in 2010. Wan (2014) found that male students performed better than female students in a high school in Sichuan. In terms of specific mathematics content areas, gender differences also existed. Ye (2011) found that male students performed better than female students in sets and simple logic, plane vectors, permutation and combination, but female students performed better than male students in trigonometry, conics, limits and differentiation, complex numbers, and analytic geometry. The inconsistency in mathematics performance between male and female students indicate that further evidence is needed.

PURPOSE OF THE STUDY

Despite the critical importance of the NHEEE test, lack of knowledge of the reliability and validity of the NHEEE test scores remain. This study sought to fill this gap in the literature of mathematics education only. The domain of mathematics was chosen out of convenience. The overall purpose of this study was to offer evidence for the reliability and validity of the NHEEE mathematics test scores.

This study addresses the following research questions: (a) Is there evidence that the NHEEE mathematics test scores can be used to measure the intended constructs? (b) Using modern measures of score reliability, is there evidence that the NHEEE mathematics test scores measure the intended constructs reliably? (c) Is there evidence that the NHEEE mathematics test scores are invariant between female and male students? and (d) Is there evidence of predictive validity of the NHEEE mathematics test scores?

METHODS

Participants

The participants were 637 (66% males and 34% females) Grade 11 students in the science concentration/track from a high school in a suburban area near Wuhan, the capital city of Hubei Province in China. This school was chosen because it was a typical suburban school in that area and represented the target population (top-tier high school students) in Wuhan. The NHEEE is designed to cover the important topics students learn in Grades 10-12. Students at lower levels were not selected because they may have forgotten what they had learned at Grade 10. As mentioned in the introduction, top-tier schools, such as this school chosen as our research site, usually accelerate the teaching pace to free some time for their students to review the contents in three rounds in Grade 12. Normally the first round is to go over the main contents chapter by chapter, the second round is to learn how to solve typical problems in each content area, and the last round is to do tens of mock examination tests. After three rounds of review in the last year, many items may become routine for many students. That is why we selected grade 11 high ability students who had learned but had not started the review process.

Procedures

The 2014 NHEEE mathematics test (National Paper I, Science Stream) was administrated, in July 2014, as an ordinary test they took every week in their routine. To simulate the context for NHEEE, the students were asked to finish the test within two hours without the use of calculators. In 2015, additional data were collected from this group of students as to their actual 2015 NHEEE mathematics test scores and the levels of the universities they were admitted to. These universities were coded as top-tier, second-tier, and third-tier, respectively.

Instrument

The 2014 NHEEE mathematics test (National Paper I, Science Stream) was selected for the following reasons: (1) It was set by the Ministry of Education; and therefore was often used as a model for the examination authorities in individual provinces to develop their own versions; (2) It was taken by the most number of students across different provinces. A province could decide whether they would like to use the NHEEE or develop their own provincial examinations in 2014; and (3) NHEEE will be used nationally again in most provinces of China in 2017 (The State Council, 2014).

In terms of item format, there were 12 multiple-choice items, 4 short-answer questions, and 8 open-response questions. The last three questions pertained to the elective topics in high school mathematics curriculum in China. The topics included plane geometry, parametric equation and its applications, and inequalities. Among them, a student only needed to answer one. If he/she answered more than one, only the response to the first one would be graded.

TABLE 1
Content Areas Covered in the 2014 NHEEE National Test

Content Areas	Item No.	Weighting	Subtotal
Set, numbers, and their operations			10 (7%)
Set	1	5	
Complex Number	2	5	
Algebra			59 (39%)
Odd/even functions	3	5	
Trigonometric functions	6	5	
Trigonometric value	8	5	
Algorithm (Input-Output)	7	5	
Linear programing	9	5	
Quadratic, cubic functions/equations and other power functions	11	5	
Binomial theorem	13	5	
Arithmetic/Geometric and other kinds of number sequences	17	12	
Differentiation and Tangent	21	12	
Inequality	24*	10	
Data analysis and probability			17 (11%)
Probability	5	5	
Data Analysis and Probability	18	12	
Geometry and measurement			54 (36%)
Hyperbola	4	5	
Parabola	10	5	
Three views of an object in space	12	5	
Simple logic	14	5	
Vector	15	5	
Solving triangles using sine and cosine rules	16	5	
Solid geometry	19	12	
Ellipse	20	12	
Plane geometry	22*	10	
Polar and parametric equations	23*	10	

Note. Items 22, 23, 24 were elective carrying 10 for each (7%), students needed to answer only one of them. Therefore, their scores were not counted in the subtotal.

Table 1 shows the content covered in the NHEEE in terms of four mathematical areas (i.e., set, numbers and operations, algebra, data analysis and probabilities, and geometry and measurement). Since it is common that a student may need to use his/her knowledge from several topics to solve an item, the content area showed in Table 1 represents the major topic tested in each item. The first author determined the content area for each items first; then an experienced high school mathematics teacher was invited to check whether the content area determined by the first author was appropriate. Very few discrepancies were found, and the disagreements were resolved after discussion.

The information in Table 1 showed that the weighting for the four content areas. Function and geometry are the two most fundamental and important components in high school mathematics as well as in NHEEE mathematics tests (Ren, 2001). For example, in 2000, the function topics took up about 37% of the total scores of the NHEEE mathematics test (Ren, 2001).

Scoring

As aforementioned, there are three kinds of problems, multiple-choice items, short-answer questions, and open-response questions. For multiple-choice items and short-answer questions, students' responses were scored as either "1" (correct) or "0" (wrong). For the open-response questions, a 0-4 scoring scale was used: 4 = correct answer with an appropriate solution process; 3 = correct answer with 75% of the solution process or 100% of the solution process but with errors in computation; 2 = 50% of correct answers; 1 some (less than 50%) correct processes; and 0 = no understanding of the problem at all.

The first author discussed the scoring criteria for each item with two graduate students in education. Then the two students coded the responses from the participants separately. The percentage of agreement in scoring of all dichotomous items was greater than 98.4% and the percentage of agreement in scoring of polytomous items was 89-99%. Discrepancies were resolved through a discussion among the two students and the first author.

Data Analysis

The 19 common items and three student self-selected items were concurrently calibrated using the dichotomous Rasch model (Rasch, 1960) and Masters' Partial Credit Rasch model (PCM; Masters, 1982) in Winsteps software (Linacre, 2009). Estimating parameters for the common and student self-selected items simultaneously in a single calibration run assures that all parameter estimates are on the common scale. The student self-selected items that are not taken by a group of students are treated as not reached or missing (Lord, 1980). One item (item 13) was excluded from the current analysis because it was not correctly edited in the file that we downloaded from the internet. It is the first short-answer question, which is normally easier than item 14.

The unidimensionality of the measure was examined using Rasch Principal Components Analysis of Residuals (PCAR) and item Mean Square (MNSQ) fit values as implemented in Winsteps. The MNSQ fit values between 0.6 and 1.4 were considered reasonable (Bond & Fox, 2007). A variance of greater than 50% explained by the Rasch dimension with the additional dimension accounting for less than 5% of the unexplained variance is considered adequate (Linacre, 2009). For polytomous items, the effectiveness of the rating scale category was evaluated based on the criteria outlined in Wolfe and Smith (2007). Reliability was examined using person and item reliability and person and item separation indices. The hierarchy of item difficulties and its relationship to person abilities were examined using the person-item map. Differential item functioning (DIF) analysis was performed between male and female students. The difference in item difficulty estimates (i.e., DIF contrast) greater than 0.5 logits with $p < .05$ is considered substantial (Linacre, 2009), and further investigation is warranted.

Pearson correlation coefficients and analysis of variance (ANOVA) were used to examine the predictive validity of the NHEEE scores. The 2014 NHEEE mathematics test scores were correlated with their 2015 NHEEE mathematics test scores. The 2014 ANOVA was used to see if significant differences exist in their performance on the NHEEE mathematics test in 2014 with respect to the tiers of universities they were admitted based on their performance in 2015 NHEEE total scores.

RESULTS

Descriptive Statistics

Table 2 displays the descriptive statistics of the item-level scores. Overall, multiple choice items (Items 1-12) and short-answer questions (Items 14-16) are quite easy for the participants as indicated by the high values of item means to the items' full scores. Of the 15 dichotomous items, item 14 is the easiest item, whereas item 16 is the hardest item. About 98% of the participants answered item 14 correctly, whereas only 57% of them answered item 16 correctly. Of the five compulsory open-response items, item 17 is the easiest item and item 21 is the most difficult item. As the first open-response item, item 17 is no doubt to be the easiest item among all the open-response items. Item 21 is the last item that all the students needed to answer. It was often called "finale item" (called *Yazhou Ti 压轴题*) of a test. This seems to be the most difficult item because only one student got a rating of 3 and one student got a rating of 4. Of the three student self-selected items, item 23 is the easiest item followed by item 24 and item 22.

Of the three student self-selected items, item 23 seems to be the easiest item and the most popular choice. Item 23 is related to ellipse and line in analytic geometry, which was just learned in the past year. It is not surprising that a higher percentage of participants selected this item. Item 22 is related to geometrical proof, in particular, circle and its inscribed quadrilateral. This is a topic that was covered in junior middle school, therefore, many students might have forgotten this topic and might not feel confident to solve it. Item 24 is related to inequality, which is normally regarded as a difficult topic for high school students (Hill, 2007). Some students did not choose any of the three questions to answer, which is not surprising in mathematics test (Jiang, Hwang, & Cai, 2014).

Rasch Analysis

The Rasch model was applied to address the first research question. Results of the principal component analysis of the residuals indicated that the unidimensionality assumption was tenable. The Rasch model assumes unidimensionality, so unidimensionality was examined with Mean Square (MNSQ) item fit statistics. The MNSQ fit values between 0.6 and 1.4 were considered reasonable (Bond & Fox, 2007). The Rasch dimension explained 57% of the variance in the data.

The largest secondary dimension accounted for only 3.7% of the unexplained variance, with an eigenvalue of 1.7. The fit statistics for all items were within acceptable limits: The infit MNSQ ranged from 0.88 to 1.02; the outfit MNSQ ranged from 0.84 to 1.75.

TABLE 2
Descriptive Statistics of Item Raw Scores

Item	*n*	Item format	*M*	*SD*
Item 1	637	Multiple-choice item	.96	0.20
Item 2	637	Multiple-choice item	.92	0.27
Item 3	637	Multiple-choice item	.96	0.20
Item 4	637	Multiple-choice item	.71	0.46
Item 5	637	Multiple-choice item	.93	0.25
Item 6	637	Multiple-choice item	.82	0.39
Item 7	637	Multiple-choice item	.93	0.25
Item 8	637	Multiple-choice item	.85	0.35
Item 9	637	Multiple-choice item	.83	0.38
Item 10	637	Multiple-choice item	.81	0.39
Item 11	637	Multiple-choice item	.59	0.49
Item 12	637	Multiple-choice item	.59	0.49
Item 14	637	Short-answer item	.98	0.13
Item 15	637	Short-answer item	.84	0.36
Item 16	637	Short-answer item	.57	0.49
Item 17	637	Open-response item	2.27	1.15
Item 18	637	Open-response item	1.97	1.56
Item 19	637	Open-response item	1.31	1.05
Item 20	637	Open-response item	1.43	1.22
Item 21	637	Open-response item	.58	0.57
Item 22	130	Open-response item	1.39	1.01
Item 23	271	Open-response item	1.89	0.78
Item 24	44	Open-response item	1.59	1.39

For all polytomous items, category disordering was observed (i.e., the average measure did not increase with the category level). An examination of the Rasch category probability curves (as shown in Figure 1 indicated that a rating category of 1 in item 17 seemed underutilized. For item 19, rating categories of 2 and 3 may not have been quite distinct as shown in Figure 2. A similar pattern was observed for items 18, 20, 21, 22, 23, and 24. A rating category of 3 was underutilized for items 21, 22, 23, and 24 as shown in Figure 3. These relatively low frequency categories may cause the category disordering.

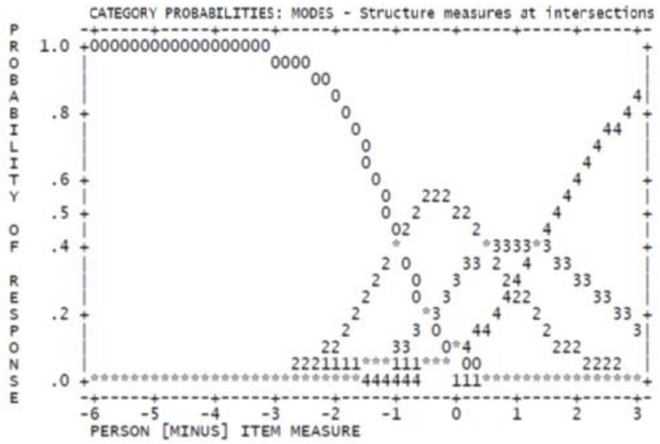


Figure 1. Category Probability Curve for Item 17

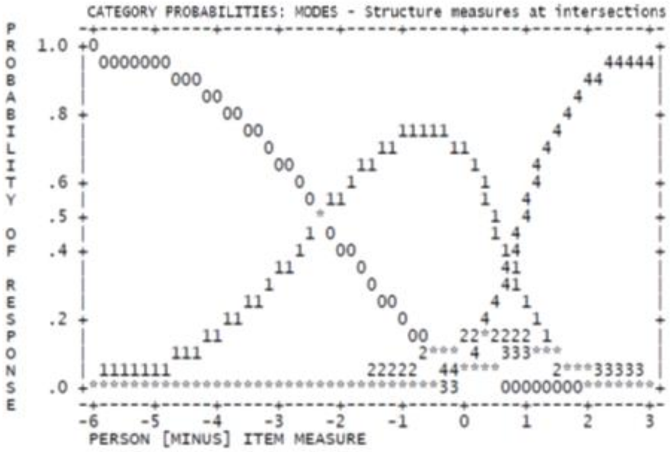


Figure 2. Category Probability Curve for Item 19

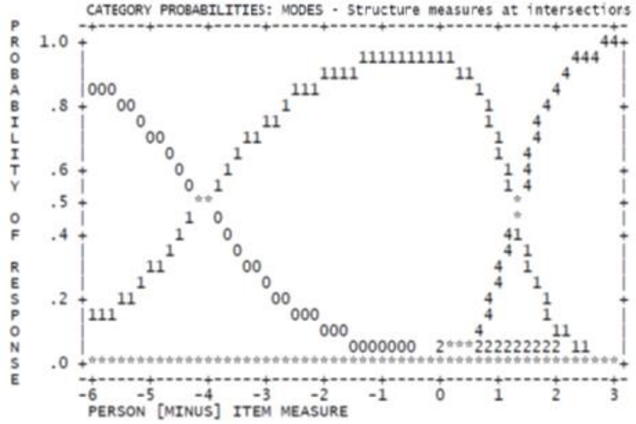


Figure 3. Category Probability Curve for Item 22

The above analyses of the rating scale structure indicated that the original five-category (01234) rating structure did not function effectively. Therefore, collapsing scale categories is warranted in order to optimize the quality of rating scale categories for valid interpretations of results. The original scale categories were collapsed with adjacent categories in different ways based on the statistical and substantive information, and then data were reanalyzed to compare alternative categories. The first analysis investigated a four-category structure (01223) which collapsed categories 2 and 3 into a single category. The second analysis investigated a three-category structure (01122) which collapsed categories 1 and 2 into a single category and collapsed categories 3 and 4 into a single category as well.

Although the four-category structure (01223) somewhat improved the distinctiveness of each category, disordered thresholds were still found for items 17, 18, 19, 22, and 24. Ratings 1 and 2 were not distinguishable from adjacent ratings. The four-category also did not improve item fit (infit MNSQ = 0.88~1.02; outfit MNSQ = 0.84 ~1.75) and reliability measures (person separation=1.40~1.58; person reliability=0.66~0.71).

The three-category structure (01122) showed noticeable improvement in the effectiveness of the rating structure. First, none of the polytomous items showed disordered thresholds. Item fit indices remain acceptable except for items 7 and 14. Outfit MNSQ values for those items were quite large (2.14 for item 7 and 4.43 for item 14). Since large outfit MNSQ values may imply the presence of unexpected rare extreme (Linacre, 2002), person fit indices were further examined to identify and remove improbable responses. The person fit analyses showed that 10 students had outfit standardized fit statistics (ZSTD) values greater than 2, indicating less compatible with the model than expected (Bond & Fox, 2007). Removing those 10 students with misfit yielded improved item fit values (infit MNSQ = 0.90~1.00; outfit MNSQ = 0.78~1.32). The removal of misfitting persons, however, did not improve the person reliability and separation measures (person separation = 1.32~1.48; person reliability = 0.63~0.69).

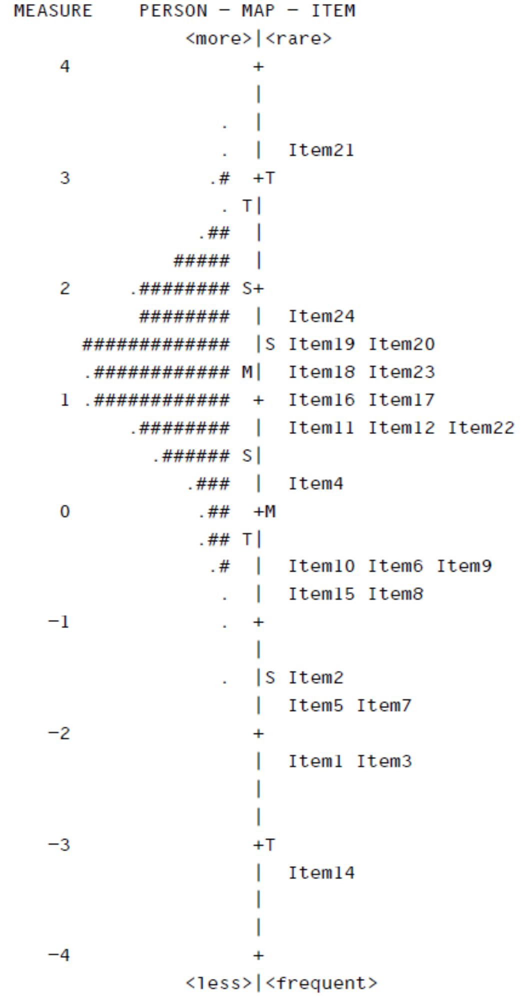
Item Difficulty and Person Ability Measures. Table 3 shows the item difficulty measures, which ranged from -3.13 to 3.32 logits (*M* = 0.00, *SD* =1.55). Item 21 was estimated to be the most difficult item. It is pertaining to a complex function with both exponential and logarithm functions. In addition, differentiation is needed to find two unknown coefficients in the first step. To prove the inequality in the second step, students need to convert the inequality into another, then to construct a new function and to find its extreme value using its monotonic property driven from its differentiation. Item 14 was estimated to be easiest item. It is pertaining to simple logic, which is an elective topic in high school mathematics.

The item location hierarchy is consistent with the expected item order (Xu, 2015). Though they are not increasing all the way from the first to the last item, in general they are increasing for the items in each format. Normally the first few multiple-choice items (all the first 10 item difficulties except item 4) and the first 1-2 short-answer items (item 14) are very easy to the students, the last 1-2 multiple-choice items (items 11-12) and the last 1-2 short-answer items (item 16) are at the level similar to the first open-response item. In addition, the difficulties of the five open-response questions are almost increasing with very tiny differences between items 19 and 20. Items 22-24 test students' abilities in solving problems pertaining to elective topics, they are set at the intermediate difficulty levels.

The person ability measures ranged from -1.57 to 3.55 logits (*M* = 1.21, *SD* = 0.78). The mean of the person ability was 1.21 logits higher than the mean of the item difficulty, suggesting that the ability levels of the sample exceeded the difficulty of the items. Logit is the natural log of odds ratios, where a value of zero means that each group has a 50% probability. In this context, a logit of 1.21 means that the person ability is much higher than the item difficulty. Figure 4 shows the person-item map where item difficulty measures (on the right side) and person ability measure (on the left side) were plotted along the latent trait being measured (the vertical line), from low at the bottom to high at the top. As shown in Figure 4, 11 items (46%) (items 1-3, 5-10, & 14-15) were too easy because their item difficulty measures are below the two standard deviation of the average ability levels of the sample. Ten items (42%) (items 11-12, 16-20, & 22-24) are at the intermediate level because their item difficulty measures are within one standard deviation of the average ability levels of the participants. Item 21 are too difficult for the sample. Although the polytomous item difficulty measures well matched most of the ability levels, dichotomous items were relatively easy. The person-item map also indicated that more difficult items are needed to measure high ability students more precisely.

TABLE 3
Item Difficulty Measures

Item No.	Item Difficulty (from most to least)
21	3.32
24	1.71
19	1.55
20	1.53
23	1.3
18	1.18
17	1.11
16	0.88
22	0.85
11	0.83
12	0.83
4	0.22
10	-0.44
6	-0.46
9	-0.56
15	-0.67
8	-0.75
2	-1.52
5	-1.65
7	-1.7
3	-2.19
1	-2.23
14	-3.13



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +
                     |
                  .  |
                  .  |  Item21
    3            .#  +T
                  . T|
                .##  |
              #####  |
    2     .######## S+
           ########  |  Item24
      #############  |S Item19 Item20
      .############ M|  Item18 Item23
    1 .############  +  Item16 Item17
          .########  |  Item11 Item12 Item22
            .###### S|
               .###  |  Item4
    0           .##  +M
                .## T|
                 .#  |  Item10 Item6 Item9
                  .  |  Item15 Item8
   -1             .  +
                     |
                  .  |S Item2
                     |  Item5 Item7
   -2                +
                     |  Item1 Item3
                     |
                     |
   -3                +T
                     |  Item14
                     |
                     |
   -4                +
               <less>|<frequent>
```

Figure 4. Person-item map
Note. Each “#” is 7 students; each “.” is 1 to 6 students.

Reliability. To address the second research question, reliability was evaluated by using person separation index, item separation index, person reliability, and item reliability provided by the Rasch analyses. The item reliability was satisfactory, as evidenced by item separation of 11.79 and item reliability of .99. The person reliability, however, was less satisfactory at 0.67 (lower bound) and 0.73 (upper bound). Person separation reliability was also low at 1.43 (lower bound) and 1.66 (upper bound).

DIF Analysis

To address the third research question, DIF analysis was conducted to examine whether the NHEEE mathematics test items functioned differentially between female and male students. The results of the gender DIF analysis showed that items 2, 5, and 6 displayed DIF, with a logit difference of 0.85, 0.79, and 0.63, respectively. A logit value of 0.50 means that each group has an equal chance. Therefore, all these values (0.85, 0.79, and 0.63) mean that these items were more difficult for female students. It will be compared with the results obtained from Ye (2011) in the discussion section.

Predictive Validity

To address the fourth research question, the 2014 NHEEE mathematics test scores were examined in relation to the 2015 NHEEE mathematics test scores and the university ranking. The Pearson correlation coefficient between the 2014 NHEEE mathematics test scores and the 2015 NHEEE mathematics test scores was .78 ($p < .001$), which means that nearly 61% of the variance in the participants' performance on 2015 NHEEE mathematics test was explained by the 2014 NHEEE mathematics test scores. ANOVA revealed statistically significant differences in the 2014 NHEEE mathematics test scores among students admitted into universities of different tiers in 2015, $F(2, 558) = 89.53$, $p < .001$. Students admitted to top-tier universities in 2015 ($M = 22.00$, $SD = 4.53$, $n = 391$) had a much higher performance on 2014 NHEEE mathematics test than students admitted to second-tier universities in 2015 ($M = 18.77$, $SD = 3.44$, $n = 101$), who had a much higher performance on 2014 NHEEE mathematics test than students admitted to third-tier universities in 2015 ($M = 16.33$, $SD = 4.31$, $n = 69$). These results provided predictive aspects of validity of the 2014 NHEEE mathematics test scores.

CONCLUSION AND DISCUSSION

This study was conducted to examine the reliability and validity evidences of the scores derived from the 2014 NHEEE mathematics test. Results of the principal component analysis of the residuals and item fit indices indicated that the unidimensionality assumption was tenable. Predictive aspects of validity were found for the 2014 NHEEE mathematics test, which suggests that these test scores are trustworthy even though the students took this test in a simulated context. However, item-level data for the 2015 NHEEE mathematics test scores were not available. Future studies should try to use the actual performance data from the NHEEE test scores rather than data collected from a simulated situation.

Analyses of the rating scale structure indicated that the original five-category rating structure did not function properly and that the three-category structure was found to be preferable. This finding potentially indicates test construction problems. As seen from the item difficulty measures, some items are relatively easy for the participants, but other items are very difficult, which means that either all students got them right or all students got them wrong. In teaching mathematics, teachers usually set up some sub-questions to help students find their solutions to difficult items, and students can get partial credits for each sub-question. Though more difficult items may increase the stress of test-takers, the use of sub-questions may alleviate the pressure and help them solve the difficult problems. This result has implications for test developers of high-stakes tests in all countries.

The hierarchy of item difficulties appeared to be consistent with the theoretical expectations, supporting evidence for construct validity for the NHEEE mathematics test. Findings indicate that the two items about set, numbers and operations are easy. There are two items about data analysis and probabilities. One is easy and the other is at the intermediate level. For the items about algebra, geometry and measurement, their difficulty levels varied from easy to the most difficult. The topic of set, numbers and operations is no longer important in high school mathematics curriculum, but the topic of data analysis and probabilities is (Chinese Ministry of Education, 2003). More items at the intermediate level about data analysis and probabilities could be included in the test. This result calls for the close match between curriculums and testing so that what is tested reflects what is taught in the classroom.

The Rasch analysis showed that the item reliability and item separation were satisfactory; however, the person reliability and person separation indices suggested a less-than desirable reliability for high-stakes tests such as the NHEEE. The low person separation reliability indicates that the exam is not sensitive enough to distinguish between low- and high-performing students. The rating categories being indistinguishable may have contributed to this finding. A more discriminating rating scale is needed to improve person reliability and person separation.

The test as a whole was somewhat easy for the participants. One plausible explanation for this finding relates to a select group of students under study who represent top-tier high school students in the suburban area of Wuhan. Although the participants are relatively high ability students, they had not started the final year review process when the study was conducted. The results can be taken as from average ability students taking the NHEEE in the next year. Eight of the 12 multiple-choice items and two of the short-answer items were relatively easy. If item 13 was counted, three of the four short-answer items would be relatively easy. For example, for item 10, more than 80% of the participants got the correct answer. These easy items probably are set to release the pressure of the students when they take the NHEEE. However, they are too easy to function well for selecting candidates for higher education. They will be easier for the participants after intensive test preparation for the NHEEE during the senior year of high school. This may also have contributed to the relatively low person reliability.

This result has significant implications for the educational policy in China as well as in many other countries implementing high-stakes tests. As mentioned in the introduction, Chinese students spent nearly all their time on the preparation for the NHEEE scarifying their time for physical activities, music and art education, and play in order to achieve higher scores on the test (Yang, 2006). However, our results suggest that the test is not sensitive enough to distinguish between low- and high-performing students because most of the test-takers got many questions correct! Although it seems to imply that Chinese government should raise the bar even higher by asking mathematics educators to add more difficult items, this is definitely not what we want. On the contrary, we urge Chinese teachers to teach less to the test and give students more time for physical activities, music, and art education as well as social and cultural events to provide a more complete and balanced educational experience for the development of the whole children (Lambert, 2015). The Chinese news media is pushing to reduce the students' load by making the NHEEE easier (Cockain, 2011; Vincent, 2015; Yuan, 2011), but the findings of this study seem to suggest that more difficult items are needed in order to measure high-ability students more precisely and to distinguish between high and low performers. Given that the current load of Chinese students is already very heavy and media coverage of the NHEEE has been heavily unfavorable, the idea of adding more difficult items to the NHEEE will not be well received by educators and parents. Therefore, instead of increasing the difficulty level of the test, we urge school administrators and teachers as well as parents to reduce the load of the students by reducing the test-preparation time and homework assignments.

As aforementioned, the majority of the dichotomous items were found to be too easy. A possible explanation is that Chinese high school students are so well prepared for the test that the items lost its function to distinguish the students' academic achievement levels. This was supported by the high average scores in mathematics of participants (117.87 out of 150) found in the study of Hu, Li, and Gan (2014). Most high school teachers give their students so many homework assignment as well as model tests to help them prepare for the NHEEE (Cockain, 2011; Vincent, 2015). The scores of NHEEE may represent how much time spent on the preparation of the test instead of the students' true mathematics knowledge and skills. We call for the reduction of student load and the teaching of the content knowledge rather than preparing for the test only. Forcing students to take a test weekly during their last year of high school is harmful not only to the students' psychological well-being but also to their physical health and problematic for the validity of test scores of the NHEEE (Cockain, 2011; Vincent, 2015; Zhang, Li, Zhang, Fan, & Li, 2016).

The DIF analysis from this study suggests that three of the 24 items were more difficult for female students, which is consistent with most previous studies with high school students (e.g., OECD, 2016; Tian & Zhu, 2014; Wan, 2014). Some other studies noted that female college students outperformed male students in mathematics at college (e.g., Qiu et al., 2009). Specifically, Item 2 is related to the computation of complex numbers. The study of Ye (2011) found that female students performed better than male students on this kind of tasks. Item 6 is related to probability, permutation and combination. This finding favoring male students is consistent with the study of Ye (2011). However, Ye (2011) did not find gender difference in function problems like Item 5. Since more and more Chinese female students could have an opportunity to study in higher education institutions, further studies are needed to investigate gender differences in their difficulties in mathematics learning.

Although these data present important new information concerning the reliability and validity evidence of the NHEEE mathematics test scores, study limitations exist. Readers should be cautious when interpreting the results from this study because this study is limited by the sample of high performing students in top-tier high schools near Wuhan. Most items in the NHEEE mathematics test were found to be too easy for these students, but this might not be the case for all high school students in China or other countries. Future studies should include students from different tiers of high schools in different regions. Although the 2014 NHEEE mathematics test scores were highly correlated with the 2015 NHEEE mathematics test scores, the 2014 test scores were obtained in a simulated situation. The students' scores might be different in a real testing context. Another limitation is that only the 2014 NHEEE mathematics test scores were used, so the results are not generalizable to the tests in other subject areas and in other calendar years. With these limitations, this study is the first that examined the aspects of reliability and validity of the NHEEE test scores in China and might be the first to examine the validity of high-stakes mathematics test in such great details as well. We hope this study will bring more research into this topic.

In summary, our findings provide preliminary reliability and validity information for the NHEEE mathematics test. Comparison of the results from collapsing the categories indicate room for the improvement of rating scales. Our results suggest that the original five categories was not as good as the collapsed three categories of the rating scale. Therefore, we need to consider the best practices of grading the polytomous items. According to Messick (1995), the validity refers to the interpretation and appropriate use of the test scores (consequential validity). The role of NHEEE plays in Chinese higher education admission process is so important that the consequences of misinterpretation of the NHEEE scores can be detrimental. The NHEEE guides instructional practices and policies in schools. We suggest that Chinese government consider examining the predictive validity with student academic scores in the junior years (Grades 10-11) and use previously administered tests as formative evaluation. The results from these formative evaluations could be used to guide the instruction and provide feedback to classroom teachers about how they can help their students to improve. This is to say that the instruction should be less driven by assessment but more driven by the standards and instructional objectives. Moreover, assessment should be more criterion-based rather than reference-based to reduce the competition among students and to reduce the anxiety of their parents and teachers. In countries that administer high-stakes testing, it is of critical importance to highlight that the reliability and validity of these test scores in all subject areas should be monitored on an ongoing basis. We cannot deny the fact that educational policy has a strong impact on the design of high-stakes tests and the classroom practices. It is pivotal to keep a high-stakes test as a measurement of students' academic knowledge and problem-solving skills rather than a measurement of the frequency of model tests the students have taken to prepare for the test.

REFERENCES

Airuishen China's University Alumni Association. (2016). Report on the National Higher Education Entrance Examination *Zhuangyuan* [In Chinese]. Retrieved from http://www.cuaa.net/cur/2016/2016gkzydc/.

Andreescu, T., Gallian, J. A., Kane, J. M., & Mertz, J. E. (2008). Cross-cultural analysis of students with exceptional talent in mathematical problem solving. *Notices of the AMS, 55*(10), 1248-1260.

Beard, J., & Marini, J. P. (2015). *Validity of the SAT for Predicting First-Year Grades: 2012 SAT Validity Sample* (College Board Statistical Report No. 2015-2). New York, NY: The College Board.

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

China News. (2016). Nine million and four hundred thousand students registered for the Chinese National Higher Education Entrance Examination in 2016 [In Chinese]. Retrieved from http://www.chinanews.com/gn/2016/06-06/7895951.shtml.

Chinese Ministry of Education. (2003). *Mathematics Curriculum Standards for High Schools (Trial version)* [In Chinese]. Beijing: People's Education Press.

Chu, R., Wang, J., Wang, Z., & Ding, Y. (2005). An analysis of the mathematics examination papers for the 2005 *Gaokao* [In Chinese]. *China Examination, 2005*(11), 23-38.

Cockain, A. (2011). Students' ambivalence toward their experiences in secondary education: Views from a group of young Chinese studying on an international foundation program in Beijing. *The China Journal, 65*, 101-118. doi:10.1086/tcj.65.25790559.

Davey, G., Lian, C. D. & Higgins, L. (2007). The university entrance examination system in China. *Journal of Further and Higher Education, 31*(4), 385-396. doi:10.1080/03098770701625761.

Ding, Y. (2008). Multi-perspective thinking on the design of mathematics items in the college entrance examination under the idea of new curriculum [In Chinese]. *China Examination, 2008*(10), 33-40.

Feng, Y. (1995). From the imperial examination to the national college entrance examination: The dynamics of political centralism in China's educational enterprise. *Journal of Contemporary China*, *4*(8), 28-56. doi:10.1080/10670569508724213.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications.* Upper Saddle River, NJ: Pearson.

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the no child left behind policy initiative. *Educational Evaluation and Policy Analysis, 29*(2), 95-114. doi:10.3102/0162373707301711.

Hu, W., Li, F., & Gan, L. (2014). Does China's national college entrance exam effectively evaluate applicants? *Frontiers of Economics in China, 9*(2), 174-182. doi:10.3868/s060-003-014-0010-7.

Hyde J. S., Fennema E., Ryan M., Frost L. A., & Hopp C. (1990). Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly*, *14*(1990), 229-324. doi:10.1111/j.1471-6402.1990.tb00022.x.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences of the United States of America, 106*(22), 8801-8807. Retrieved from http://www.pnas.org/content/106/22/8801.full.pdf.

Jiang, C., Hwang, S., & Cai, J. (2014). Chinese and Singaporean sixth-grade students' strategies for solving problems about speed. *Educational Studies in Mathematics, 87*(1), 27-50. doi:10.1007/s10649-014-9559-x.

Lambert, R. G. (2015). Student perceptions of the Chinese national college entrance examination system. In C. Wang, W. Ma, & C. Martin (Eds.), *Chinese education from the perspectives of American educators: Lessons learned from study-abroad experiences* (pp. 81-99). Charlotte, NC: Information Age Publishing.

Linacre J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009). Winsteps (Version 3.68.0) [Computer Software]. Chicago, IL: Winsteps.com.

China Education Yearbook Editorial Board. (2014). *China education yearbook 2013* [In Chinese]. Beijing: People's Education Press.

Liu, H., & Wu, Q. (2006). Consequences of college entrance exams in China and the reform challenges. *KEDI Journal of Educational Policy, 3*(1), 7-21.

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

Madaus, G. F. (1991). The effects of important tests on students: Implications for a national examination system. *The Phi Delta Kappan, 73*(3), 226-231.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037//0003-066x.50.9.741.

OECD. (2013). *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science (Volume I)*, PISA, OECD Publishing. doi:10.1787/9789264201118-en.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and equity in education,* PISA, OECD Publishing, Paris. doi:10.1787/9789264266490-en.

Olsen, A. (2009). *The Gaokao: Research on China's National College Entrance Examination*. Australian Education International (AEI).

Patterson, B. F., & Mattern, K. D. (2011). *Validity of the SAT for predicting first-year grades: 2008 SAT Validity Sample* (College Board Statistical Report No. 2011-5). New York, NY: The College Board.

Patterson, B. F., & Mattern, K. D. (2012). *Validity of the SAT for predicting first-year grades: 2009 SAT Validity Sample* (College Board Statistical Report No. 2012-2). New York, NY: The College Board.

Patterson, B. F., & Mattern, K. D. (2013a). *Validity of the SAT for predicting first-year grades: 2010 SAT Validity Sample* (College Board Statistical Report No. 2013-2). New York, NY: The College Board.

Patterson, B. F., & Mattern, K. D. (2013b). *Validity of the SAT for predicting first-year grades: 2011 SAT Validity Sample* (College Board Statistical Report No. 2013-3). New York, NY: The College Board.

Qiu, X., Chen, L., & Xiao, P. (2009). An analysis of the state of undergraduate mathematics learning [In Chinese]. *Journal of Fujian Normal University (Natural Science Edition), 25*(2), 119-124.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Ren, Z. (2001). Analysis report on the 2000 national higher education entrance mathematics examination papers [In Chinese]. *China Examinations*, *2001*(3), 11-13.

Schultz, A. (2015, October 30). U.S. colleges put China's Gaokao to the test: High marks in China's grueling exam, plus good English skills, can win admission into some schools. *Barron's Asia*. Retrieved from http://www.barrons.com/articles/u-s-colleges-put-chinas-gaokao-to-the-test-1446188818.

Song, W., & Zhang, S. (2017). Comparative analysis of physical health of male and female students of local normal universities: Taking Jilin Normal University as an example [In Chinese]. *Journal of Educational Institute of Jilin Province, 433*, 181-183.

The State Council. (2014). *Implementation Recommendations for Deepening the Reform of Examination and Admission System* [In Chinese]. Retrieved from http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/moe_1778/201409/174543.html.

Tian, J., & Zhu, Q. (2014). Gender differences research and enlightenment for mathematics learning based on College Entrance Examination data [In Chinese]. *China Examinations*, *2014*(7), 19-22.

Vincent, D. (2015, June 11). China's pressure-cooker schools. Retrieved from http://www.bbc.com/capital/story/20150611-chinas-parent-army.

Wan, Y. (2014). Factors causing gender gap in mathematics performance and teaching strategies [In Chinese]. *Reading, Writing, and Calculating*, *2014*(12), 311.

Wang, L. (2008). Rasch measurement principles and implementation in the entrance examination to higher education evaluation [In Chinese]. *China Examinations, 2008*(1), 32-39.

Wang, X. B. (2006). An introduction to the system and culture of the college entrance examination of China. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2006-28-introduction-system-culture-college-exams-china.pdf.

Woessmann, L. (2001). Why students in some countries do better: International evidence on the importance of education policy. *Education Matters*, *1*(2), 67-74.

Wolfe, E. W., & Smith, Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement, 8*(2), 204-234.

Wu, G. (2007). *Study on the Issue of Validity of National College Entrance Examination* [In Chinese]. Unpublished doctoral dissertation, Xiamen University, China.

Wu, M., & Hornsby, D. (2012). Inappropriate uses of NAPLAN results. *Practically Primary, 17*(3), 16-17.

Wuhan Wanbao (2012). The trends go on that more female students than male students are studying in higher education institutions in China [In Chinese]. Retrieved from http://edu.qq.com/a/20120909/000016.htm.

Xiong, B. (2016). The admission system of key universities in China [In Chinese]. Retrieved from http://finance.sina.com.cn/roll/2016-05-19/doc-ifxsktkr5729052.shtml

Xu, G. (2015). Setting questions for Gaokao, a long-lasting split is always followed by a reunion, and vice versa: The change brought to the 2016 Gaokao in Guangdong [In Chinese]. *Studies in Middle School Math, 2015*(9), 21-24.

Yang, D. (2006). Pursuing harmony and fairness in education [In Chinese]. *Chinese Education & Society, 39*(6), 3-44.

Yang, M., He, X., & Ning, L. (2010). Survey on mathematics study of university students [In Chinese]. *Journal of Mathematics Education*, *19*(6), 56-59.

Yang, X. (2007). *Historical review of national higher education entrance examination in China* [In Chinese]. Hubei: Hubei Changjiang Publication Limited and Hubei People's Press.

Ye, H. (2011). Gender differences in mathematics performance in National Entrance Examination [In Chinese]. *Modern Education Science, 2011*(6), 41-42, 87.

Yu, L., & Suen, H. K. (2005). Historical and contemporary exam-driven education fever in China. *KEDI Journal of Educational Policy*, *2*(1), 17-33.

Yuan, G. (2011). Promote the scientific development of China's education cause [In Chinese]. Retrieved from http://www.edu.cn/zong_he_news_465/20110105/t20110105_566588.shtml.

Zang, H., & Sun, H. (2014). *China's Yearbook 2014 of National Entrance Examinations to College: Mathematics* [in Chinese]. Neimenggu: Neimenggu Publication Limited.

Zhang, D., Li, S., & Tang, R. (2004). The "two basics": Mathematics teaching and learning in Mainland China. In L. Fan, N. Y. Wong, J. Cai, & S. Li (Eds.), *How Chinese learn mathematics: Perspective from insiders* (pp.189-207). Singapore: World Scientific Publishing Co. Pte. Ltd.

Zhang, H. (2015, June 26). Italy, France accept *Gaokao* scores. *China Daily*. Retrieved from http://europe.chinadaily.com.cn/china/2015-06/26/content_21110695.htm.

Zhang, H. (1988). Psychological measurement in China. *International Journal of Psychology, 23*(1-6), 101-117.

Zhang, S., Li, X., Zhang, T., Fan, Y., & Li, Y. (2016). The experiences of high school students with pulmonary tuberculosis in China: A qualitative study. *BMC Infectious Diseases, 16*(1), 758-765. doi:10.1186/s12879-016-2077-y.