Special Issue:
Interfacing Between Test Developers and Their Constituents

JOURNAL OF APPLIED MEASUREMENT

Volume 24, Number 1/2 2023

Visit our website at https://jamntnu.net/

*This double issue is for Volume 24 (2023), printed and published in 2024.

JOURNAL OF APPLIED MEASUREMENT, *24*(1/2), 58-87

Examining Equivalence of Three Versions of Mathematics Tests in China's National College Admission Examination Using a Single Group Design

Chunlian Jiang
University of Macau

Stella Yun Kim
University of North Carolina at Charlotte

Chuang Wang
University of North Carolina at Charlotte

Jincai Wang
Soochow University

The National College Admission Examination, also known as *Gaokao*, is the most competitive examination in China because students' scores obtained in Gaokao are used as the only criterion to screen applicants for college admission. Chinese students' scores in Gaokao are also accepted by many universities in the world. The one-syllabus-multiple-tests practice has been implemented since 1985, but not much has been explored as to the extent to which multiple tests are equivalent. This study attempts to examine the equivalence of three versions of Gaokao mathematics tests and to illustrate the methodological procedure using a single group design with an item response theory (IRT) approach. The results indicated that the three versions were comparable in terms of content coverage; however, most items were found to be easy for the students so more challenging items are suggested to be included for distinguishing students with average and high mathematics competencies. Some differences were also noted in terms of differential item functioning analysis and the factor structure.

Keywords: test concordance, mathematics examinations, item response theory, single group design, Gaokao

---

Requests for reprints should be sent to Chunlian Jiang, Faculty of Education, University of Macau, Avenida da Universidade, Taipa, Macau, China; email: cljiang@um.edu.mo
This project was sponsored in part by the University of Macau (UM) Multi-Year Research Grant. We are grateful to UM for the support. Opinions expressed herein are those of the authors and do not necessarily represent the views of UM.

Students from Asian countries often perform better than their peers from Western countries in international comparisons (Mullis et al., 2016; Organization for Economic Cooperation and Development [OECD], 2016), high-stakes exams might be one of the factors contributing to their relatively better performance (Woessmann, 2001). In the top-performing Asian countries, there are all high-stakes examinations. For example, Singaporean students have to take the Primary School Leaving Examination (PSLE) after completing their six-year primary education, the General Education Certificate (GCE) "O-Level" examination at the graduation of Grade 10, and the GCE "A-Level" examination for higher education opportunities at the graduation of Grade 12. Their performance in the three examinations is all used for placement and stratification, and thus the examinations are generally taken as high-stakes examinations (Gregory & Clarke, 2003; Tan, 2017). Similarly, in China, students take the High School Admission Examination (also known as *Zhongkao*) at the end of Grade 9 and the National College Admission Examination (NCAE; commonly known as Gaokao) at the end of Grade 12. Their scores in Zhongkao are used to screen them for admission to either high schools or vocational schools, and their scores in Gaokao are used as the only measure for the admission decision of higher education opportunities. Different from the graduation examination, which is named as "General Academic Achievement Test for High School Students" (普通高中学业水平考试 / 普通高中學業水準考試), NCAE is used primarily for selecting candidates into higher education, particular programs, and their career path (Davey et al., 2007; Lambert, 2015; Reshetar & Pitts, 2020; Wang, 2006). Therefore, it is highly competitive because the admission decision is made based on the rank order of test-takers' scores relative to their cohort (Wang, 2006).

Previous research suggests that there are both pros and cons to high-stakes examinations like NCAE (Gregory & Clarke, 2003; Gu et al., 2017; Zhao, 2020). NCAE has been criticized for the inappropriate use of the scores which cannot predict the future development of the so-called high achievers and for the fact that it ruins students' interests in mathematics, the joy of acquiring new knowledge, and their independent thinking in mathematics learning (Zhao, 2020). It is also criticized for the undesirable "backwash" or "trickle-down" on classwork and study of students at lower grade levels (Gu et al., 2017). On the other hand, some students who took Gaokao thought that the Gaokao journey strengthened their abilities to tackle pressure, make self-adjustment, manage their time well, and so forth (Gu et al., 2017). However, NCAE cannot be replaced with any other assessment for its fairness and efficiency in the Chinese context (Gu et al., 2017; Wang, 2010). Nowadays, Chinese families invest a lot of time and money in their children in preparation for NCAE. Despite the fact that NCAE plays a central role in Chinese students' education and future career path, the quality of the tests in subject areas has not been studied thoroughly (Wang, 2010). This study, therefore, aims to fill this gap by examining the equivalence of three versions of mathematics tests in NCAE.

Theoretical Framework and Related Literature

National College Admission Examination (NCAE)

NCAE is the result of the evolvement of the *Keju* Examination, which originated in the Sui dynasty (581–619; Yang, 2021) and ended in the Qing dynasty (1616–1911) in China to select talented officials to serve the country (Feng, 1995; Yu & Suen, 2005; Zhang, 1988). Nowadays, NCAE scores play a critical role in

college admission, the assessment of the job performance of teachers and administrators as well as the ranking of high schools (Lambert, 2015). Recently, NCAE scores started to serve as a gold standard for accepting Chinese high school graduates for undergraduate programs in universities in 20 countries, including Australia, the United States, and the United Kingdom (Olsen, 2009; People's Daily Online, 2019; Schultz, 2015; Zhang, 2015). Although the college admission rate in China has dramatically increased in the past two decades from 27.3% in 1990 to 87% in 2012 (China Education Yearbook Editorial Board, 2014), and 90% in 2020 (China Education Online, 2021). Graduates from top-tier universities are at a great advantage in the job market, which reinforces the important role of NCAE scores (Davey et al., 2007; Liu & Wu, 2006).

NCAE has been revised several times, and more freedom has been given to local provinces to develop their own tests to replace the National test developed by the National Education Examinations Authority (NEEA), Ministry of Education of People's Republic of China (MOEPRC; Yang, 2007). In 1985, the first local test was developed in Shanghai for college admission of Shanghai students. Beijing, Guangdong, and Henan followed and used their own tests. Since 2004, MOEPRC allowed more provinces to have their own tests as long as they were in line with the general guidelines. Till 2014, more than 20 sets of mathematics tests of NCAE were adopted, and then questions were immediately raised about the validity and equivalence of these tests with empirical studies. In 2015 many provinces abandoned their own tests to avoid disputes and adopted the National test again. Therefore, 2014 is the year when the number of mathematics tests was at a maximum (Liu, 2015), which is one of the reasons to select test papers used in this year. In 2022, only three municipalities (Beijing, Shanghai, and Tianjin) and Zhejiang Province used their local tests, and the other 27 municipalities and provinces used the National tests developed by NEEA. The decision to adopt a local test or the National one was made with little research that examines the reliability, validity, equivalence, and even basic psychometric properties of these test scores (Hu et al., 2014; Jiang et al., 2019; Reshetar & Pitts, 2020; Wang, 2006; Wu, 2007). The current study aims to fill this gap by examining the equivalence of three tests (i.e., the National, Hunan, and Jiangsu tests) with a focus on mathematics. Why the three were chosen will be explained in the Methods section. The study focused on the subject of mathematics because mathematics is one of the three core subject tests that all students have to take. The results from the current study will provide useful information for educators, policymakers, and researchers in terms of how to demonstrate equivalence of different test forms.

NCAE Mathematics Tests

The mathematics test in NCAE is used to test candidates' mathematics foundations necessary for their future study and their learning abilities with selection as the primary mission of mathematics tests in NCAE (NEEA, 2016). Of course, the mathematics test in the NCAE also serves other purposes, such as enhancing social fairness, promoting students' core competencies, and playing its guiding role in education reform in high schools (NEEA, 2016). The NCAE mathematics test is intended to measure students' basic mathematics knowledge and skills and to measure their understanding and application of basic mathematics ideas and methods (Chu et al., 2005). Basic mathematics knowledge includes knowledge of functions, equations, inequalities, conics, vectors, and trigonometric functions (Chu et al., 2005), while basic mathematics skills include spatial imagination, abstract thinking, reasoning skills, computation skills, data handling skills, application skills, and creative skills (NEEA, 2019a). Basic mathematics ideas include functional and equivalence ideas, integral ideas of numbers and graphs, classifying and combination ideas, transformation ideas, ideas about special cases and their generalizations, finite versus infinite ideas, and ideas about certainty and uncertainty.

Basic mathematics methods include cutting and patching, proof by contradiction, substitution methods, and so forth (Chu et al., 2005). One item may be related to more than one content area and measures more than one skill. For example, the item "Given that $A = \{x \mid x^2 - 2x-3 \geq 0\}$, $B = \{x \mid -2 \leq x < 2\}$, *find* $A \cap B$" measures the understanding of the symbols, the ability to solve quadratic inequality, and the skills to draw a number line to find the intersection of the two sets. Therefore, logical thinking, computational, and spatial imaginary skills are required to solve it correctly. Therefore, we shall only examine the main content areas the test items were designed to measure and then examine their relative difficulties with respect to the main thinking skills tested.

The mathematics test in NCAE is developed based on the examination syllabus for Gaokao (NEEA, 2014, 2019a) and the mathematics curriculum standards for high schools (NEEA, 2022). The high school (Grades 10–12) mathematics curriculum can be classified into compulsory, required elective, and free elective courses (MOEPRC, 2003). Before 2017, there were 16 elective courses in college mathematics (e.g., Selected Lectures on the History of Mathematics, Information Security and Secret Code, Spherical Geometry); however, only two to four topics were listed in the examination syllabus. For example, only two topics were included in the National examination syllabus (NEEA, 2014), and four were included in the Jiangsu examination syllabus (Jiangsu Education Examination Authority, 2013). In the 2017 version of the curriculum standards for high school mathematics, elective courses were regrouped into A, B, C, D, and E categories (MOEPRC, 2017). The courses in Categories A to D are for students in the science stream, social studies stream, arts and humanity stream, and sports and arts stream, respectively. Category E includes courses that broaden students' horizons, courses with local characteristics, and advanced placement courses in college mathematics. However, the elective courses have not been tested in the mathematics test of NCAE since 2021 (MOEPRC, 2019; NEEA, 2021). What is tested in Gaokao will be what is to be taught and what is to be learned in high schools, which was also criticized by mathematicians (Yang, 2007; Zhao, 2020).

The mathematics test in NCAE was developed by an item development panel that consisted of seven to ten teachers with five to seven professors from higher education institutions and two to three from high schools. Three professors were selected to form the core group and acted as the principal team leader, the deputy team leader, and the secretary, respectively. The professors were responsible for developing examination questions, and the high school teachers were involved in the pilot test as test takers and were charged to screen the items to see whether they were assessing the content areas listed in the high school curriculum standards. The item development was conducted as follows:

1. The professors acquainted themselves with the types of items to be included and their difficulty levels by studying the curriculum standards, examination syllabus, mathematics textbooks, and the information (e.g., psychometric properties) about the tests used in the previous years.
2. They designed a two-way specification table for the "current" year to which each item was assigned according to two dimensions: the content areas to be covered and the cognitive level of the item in the corresponding content areas.
3. They constructed items. Generally, the professors in the core group were responsible for the most important items (i.e., the last two short-answer questions and the last three open-response items), and high school teachers and other professors were responsible for the simple items (i.e., multiple-choice items, simple short-answer items, and simple open-response items). The items developed by the core group had to be original, and the rest could be

reformulated from items included in the textbooks and tests administered in previous years. Also, at least one item had to be reformulated from textbook problems so that instruction in high schools could be aligned with the curriculum and textbooks. The item writing task assignment is not fixed; they were allowed to make appropriate adjustments whenever necessary based on their progress. Those who finished earlier could help others whose progress was lagged behind. This procedure took approximately 3–5 days but no more than a week. The items set at the end of this stage were taken as the first draft.

4. All the teachers worked together to polish every item in the first draft, starting from difficult to easy items, because difficult items were more crucial. This took another 3–5 days, and the test developed at the end of this stage was considered as the second draft.
5. The high school teachers were invited to take the test with the secretary playing the role of the invigilator. Once the teachers finished taking the test, the principal and the deputy team leaders scored their responses to the test, which allowed them to gain an understanding of the difficulty levels of the items in the second draft. Then the high school teachers shared their reflections on their problem-solving process and provided their comments on each item. The whole team, except the secretary, worked together to revise the test and to adjust the item difficulty levels and the wording. After several rounds of revision, they came up with the third draft.
6. The whole team worked closely to carefully check every item for grammatical errors and typos. During this process, one member read each item loudly, and other members listened to ensure that every item was error-free. When they found any ambiguity, they made an appropriate revision. After that, every member was charged to read the whole test three times and to carefully check for any errors in the print, including punctuation marks. This process continued until they could not find any errors, and then they submitted the final version to the respective examination authority after having it signed by all members. The whole process took about 3 weeks. After that, they started to work on a supplementary test paper just in case the first test paper was leaked. For security reasons, even after finishing their work, they had to stay in the secret location without outside contact until the test was held. Over the years, the format of Gaokao problems has been fixed; that is, similar numbers and similar types of problems were included in NCAE mathematics tests in a similar order (Zhao, 2020).

Research on NCAE Mathematics Tests

Although mathematics is one of the core subjects in NCAE, research on the NCAE mathematics tests is limited and contradictory (Hu et al., 2014; Wu, 2007). The study of Hu et al. (2014) found that students' performance on the NCAE was significantly correlated with their college grade point average. However, Wu's (2007) study revealed that the correlation coefficients between students' mathematics test scores in NCAE and their overall academic achievement throughout the four-year university study were low in all eight universities and even negative in four universities. The inconsistencies between a limited number of studies warrant further study to draw valid conclusions about the validity of NCAE scores (Wang, 2008).

Most research on the NCAE mathematics tests in China used content analysis to investigate the difficulty levels or cognitive levels of the items included. One group of researchers (Han et al., 2022; Li & Shi, 2020;

Wu & Zhang, 2018; Zhang et al., 2016; Zhang & Zhou, 2020) used the comprehensive difficulty model developed by Bao (2002), which included five dimensions (e.g., investigation, contexts, computation, reasoning, and topic coverage). Wu and Zhang (2018) extended Bao's (2002) model by adding two more dimensions (with/without parameters like θ in the equation $\begin{cases} x = \sqrt{3}cos\,\theta \\ y = sin\,\theta \end{cases}$ [θ is a parameter] which defines a curve, directions of thinking) and used it to compare the difficulty levels of items included in NCAE mathematics tests in both China and Korea in 2014–2016. They found that the comprehensive difficulty level of the NCAE mathematics test in China is higher than that in Korea. Another group (Ai & Zhou, 2017; Chen et al., 2022; Wang & Zhou, 2022; Yu, 2022) used the structure of observed learning outcome (SOLO) taxonomy theory proposed by Biggs and Collis (1982). For example, Ai and Zhou (2017) used the SOLO taxonomy to analyze items included in mathematics tests in NCAE 2016 and found that most of the items were at the multi-structural and relational levels, and the total score of items at these two levels accounted for over 80% of the whole test. Very few items were at the extended abstract level, and their total scores accounted for less than 10% of the whole test. These studies did not use the empirical data collected from students or psychometric approaches to investigate the difficulty levels of items included in mathematics tests in NCAE.

In the study of Jiang et al. (2019), the reliability and validity of the NCAE mathematics test scores were examined. Results from the item response theory (IRT) analysis indicated that the mathematics test was not sensitive enough to distinguish between low- and high-performing students. Since the student workload is already high, they suggested teaching the content knowledge rather than enhancing students' test-wiseness. The study reported the results of the National test in NCAE 2014, whereas the current study compared the same National test with two tests developed by local provinces to see how equivalent they are. This kind of research will inform us about how to improve the test design to attain a better reliability and separation index, and it will provide information to policymakers in China regarding whether or not to continue allowing local provinces to develop their own college admission tests.

Purpose of the Study

The current study was intended to address the following research question: To what extent are the three mathematics tests in NCAE equivalent?

This study selected three mathematics tests used in NCAE 2014: the National, Hunan, and Jiangsu tests. The National test for the science stream was selected for the following reasons: (a) It was set by NEAA and was often used as a model for individual provinces to develop their own versions, and (b) it was taken by the largest number of students across different provinces. A province could decide whether to use the national one or develop its own provincial tests (The State Council, 2014). Only the Hunan and Jiangsu tests were chosen because other provincial mathematics tests are very similar to the National test, and it may be meaningless to include more in the study. Another reason for selecting the Hunan test is that it is similar to the National test, but the computations involved in it are much more complicated than the National test, whereas the reason for selecting the Jiangsu test is that its format is unique in that it did not include any multiple-choice items, but included short-answer questions instead. The National and Hunan tests were chosen as exemplars of those similar tests. The examination of the equivalence of the three tests will provide us with a relatively "full" picture of the mathematics tests in NCAE 2014.

To address this research question, a single group design, where the same examinees take two different test forms, was employed (Kolen & Brennan, 2014). In particular, two groups of students were involved in the current study, one group taking the National test and the Hunan

test and the other group taking the National test and the Jiangsu test. The primary benefit of the design is that differences in scores between the two forms are entirely attributable to differences in the difficulty of the two forms because the examinees' achievement level remains the same over the two test administrations. Several studies (Nguyen, 2019; Riddle, 2005) have noted the importance of establishing equivalence across various local tests to compare students from different states, regions, or countries. However, there is a challenge associated with establishing the link across tests. To achieve this goal, one of the following conditions has to be met: (1) two (or more) tests are given to a single group of examinees, or (2) two tests are given to equivalent groups of examinees. The current study used the former design, which is less frequent in practice as it is challenging for the same group of students to complete both tests. The current study, thus, is one of the few that employed this particular design. Another advantage of this particular design includes a reduced required sample size.

Establishing equivalence of the three tests possibly falls into the category of "concordance" defined by Mislevy (1992) and Lim (1993). Concordance, or "moderation" according to Kolen and Brennan (2014), is considered when (a) inferences made from test scores are the same across tests, (b) constructs to be assessed are similar, (c) populations to which inferences are made are similar, and (d) measurement characteristics (e.g., test specifications) are dissimilar. One of the most frequently cited examples of concordance is perhaps the concordance relationships between ACT scores (ranging 1–36) and SAT I Verbal-plus-Mathematics scores (ranging 400–1600; Dorans et al., 1997).

The equivalence of the three tests was examined in terms of the content areas covered and their psychometric characteristics. This may also help to provide further useful information to policymakers for the improvement of NCAE mathematics tests.

Method

This section is divided into two parts: data collection methods and data analysis methods.

Data Collection

Participants

Data were collected from two groups of Grade 11 high school students in the science stream in the summer of 2014. One group of the examinees (n = 657) was from a high school in Hubei Province, and the other group (n = 524) was from a high school in Jiangsu Province. The two particular schools were selected because they are both top-tier schools where an accelerated instructional pace was adopted to provide students with sufficient time to review the content covered in NCAE mathematics tests in their Grade 12. For those top-tier schools, the content that is supposed to be covered in Grade 12 is taught in Grade 11 in advance, and then they usually take numerous mock examination tests in Grade 12 so that they are readily prepared for NCAE. Therefore, students were deemed qualified to take the NCAE even though the examinations were developed for Grade 12 students. Although they were from top-tier schools, they had not started their final year of study, of which the main purpose was to review the content areas that might be included in NCAE and to take mock examinations of NCAE (Zhao, 2020). They are potential candidates for the first tier of universities and are appropriate to participate as a convenience sample of this study.

Instruments

Three mathematics tests compared in this study were National, Hunan, and Jiangsu mathematics tests. The National test was composed of 12 multiple-choice (MC) items, four short-answer questions, and eight open-response items, which led to 24 items in total. Among the eight open-response items, the last three items were for the elective topics from which the students were asked to choose only one of them to answer. The Hunan test consisted

Table 1

Item Weight

National test		Hunan test		Jiangsu test	
Item	Weight	Item	Weight	Item	Weight
N1-N12[a]	5 marks	H1-H10[a]	5 marks	J1-J14[b]	5 marks
N13-N16[b]	5 marks	H11-H16[b]	5 marks	J15-J17[c]	14 marks
N17-N21[c]	12 marks	H17-H19[c]	12 marks	J18-J20[c]	16 marks
N22-N23[c]	10 marks	H20-H22[c]	13 marks	J21-23[c]	10 marks

Note. (a) Multiple-choice items, (b) Short-answer questions, (c) Open-response items.

of 10 MC items, six short-answer items, and six open-response items, leading to 22 items in total. For the Hunan test, the first three short-answer items were set for elective topics for which students were required to choose two to answer. The Jiangsu test consisted of two sections. The first section included 14 short-answer items and six open-response items. The second section included three open-response items, among which the first included four sub-items covering elective topics. Students were required to select two of the four elective sub-items to answer. The weights for items in the three tests are summarized in Table 1.

The supplementary section of the Jiangsu test and the elective items on the National test were not administered to the sample from Jiangsu because their mathematics teachers did not want them to "waste" time on these items before the review process had started.

Procedures

Besides the National test, the Hubei students were administered the Hunan test, and the Jiangsu students were administered the Jiangsu test. The administration of these tests mimicked the operational setting of NCAE. In particular, 2 hours were allowed for the National and Hunan tests and the first section of the Jiangsu test. The time interval between the administrations of the two tests was approximately one week for both groups. The administration of the National test to both groups allows us to fix the ability scale across the three tests and compare their item parameter estimates from IRT analysis.

Data Analysis

Scoring and Data Manipulation

For the MC and short-answer items, students earned 1 mark for a correct response and 0 for an incorrect response. The open-response items were scored from 0 to 4 marks, with 0 indicating "blank or unacceptable solution" and 4 representing "correct answer with appropriate solution process." For each open-response item, the first author presented all the possible solutions first, then developed a scoring criterion with key working steps reached for receiving a 2/3/4 marks through a discussion with two graduate students who graded the students' responses individually item by item. The graduate students had obtained their bachelor's degree in mathematics education and were pursuing their master's degrees in mathematics education. It took them about one week to finish scoring all the participants' responses. The percentage of agreement in scores on the dichotomous items was greater than 98.4%, and the percentage of agreement in scores on the polytomous items was 88–99%. Any disagreement was resolved through a discussion among the graders and the first author to reach an agreement on the final scores.

For the National test, one item was omitted from the analysis as it was mistyped in the test

booklet. Operationally, the selected tests use pre-defined item weights to form final scores, having each content area's contribution to the final score aligned with the test blueprint. After applying the item weights shown in Table 1, the full mark was 150 marks for the National and Hunan tests and 160 marks for the Jiangsu test.

The following six distinct procedures were conducted to explore the equivalence of the three tests: content analysis, item analysis, item/test information, reliability analysis, correlation analysis, and differential item functioning (DIF) analysis. In addition to these analyses, one additional calibration was executed to examine the impact of not using the single group design. That is, each form was calibrated separately without considering the link (i.e., the National test) among them, and results were compared between concurrent calibration under the single group design and separate calibration.

Content Analysis

The content analysis is crucial in assessing the equivalence of tests. It was carried out in two steps. The first step was to examine test specifications or test blueprints. In particular, we examined the *Examination Syllabus for the NCAE 2014 (Science Stream)* issued by NEEA (2014), the *Description of Hunan Version of the NCAE 2014* issued by Hunan Education Examination Authority (2014), and the *Description of Jiangsu Version of the NCAE 2014* issued by Jiangsu Education Examination Authority (2013). They were all developed according to the *Mathematics Curriculum Standards for High Schools* (Trial Version; MOEPRC, 2003). The examination syllabus issued by NEEA was also used as an important reference for the development of the Hunan and Jiangsu versions.

The second step was to compare the three tests in terms of content areas through an evaluation of each item in the test. The first author determined the content areas of the items in each test, and then an experienced mathematics teacher was invited to check whether the content areas determined were appropriate. Since it is common to see that a student capitalizes on his/her knowledge from various content areas to solve one item, only the primary content area was identified for each item.

In addition, we also examined the examination syllabi for 2014–2019 issued by NEEA and found that there were no significant changes made during the period. In the autumn of 2019, NEEA released *Assessment System for China's Gaokao* as a general guideline on the test development of all the Gaokao subjects. The core functions of Gaokao are: (a) To strengthen moral education and cultivate the people, (b) to facilitate the selection of the talented, and (c) to serve as a guide for teaching and learning practice in schools (NEEA, 2019b). The contents to be covered in the tests include (a) core values, (b) subject-based competencies, (c) key skills, and (d) essential knowledge (NEEA, 2019b). The main change in the mathematics test in Gaokao 2020 is that more items related to real-world contexts (e.g., COVID-19) were included. In addition, two new categories of items were added (NEEA, 2020, 2022). One category is multiple-choice items with multiple responses, where two or more of the options are keyed as correct answers. The other category is ill-structured items, where students were required to select one out of three propositions so that it can be combined with the given propositions, and finally, students were required to determine whether there exists such a triangle based on the proposition selected. Different selection leads to different answers. In addition, no items on the elective topics were included. We have collected a set of data for the 2021 test and shall go on with a similar psychometric analysis as Jiang et al. (2019). In the three tests analyzed in the current study, there were no items in the two categories.

Item Analysis

Item characteristics, including item difficulty and item discrimination, were examined using IRT. The one-parameter logistic (1PL) IRT model was fit to the

dichotomously-scored items, and the graded response model (GR; Samejima, 1969) was fit to the polytomously-scored items. Concurrent calibration allowed the estimation of all the item parameters from the three tests simultaneously. Combined data from the two groups of examinees were used as input for flexMIRT version 3.5 (Cai, 2018), and items not taken by a particular group were regarded as not reached or missing (Lord, 1980). The items on the National test served as links to put the ability and item parameter estimates on the same scale. In flexMIRT, one of the group means is, by default, set to 0, and its standard deviation to 1. The group from Jiangsu was arbitrarily chosen to serve as the base group, and the mean and the standard deviation of the Hubei students were freely estimated.

Item/Test Information Function

The item information function (IIF) in IRT shows the amount of precision that each item provides at different ability scores (Lord, 1980). The test information function (TIF), as a sum of the IIFs across all the items in the test, represents the amount of precision provided by a test at different ability scores. Item and test information functions allow us to visually inspect where an item or a test provides maximum information about an examinee's ability along the ability scale. The ability score is on a different scale from the observed score, ranging from $-\infty$ to $+\infty$, and two scores usually are non-linearly related. Given the item parameter estimates, a particular ability score can be transformed into its corresponding observed score. Both IIFs and TIFs were examined for the three tests to determine if the tests provided similar information across the ability scores.

Reliability Analysis

The reliability coefficients were compared across the tests using both Cronbach's alpha (α) internal consistency reliability coefficient and IRT reliability coefficients. Two items from the National test were excluded from the reliability analysis due to the huge volume of missing responses to them. Eliminating the two items from the reliability analysis could potentially reduce the reliability coefficient due to the shortened test length.

Correlation Analysis

The correlations between test scores on the National and province-developed tests were computed. In addition to the Pearson correlation values, the disattenuated correlation coefficient, which corrects for attenuation due to the unreliability of measurement, was also considered. The correlation analysis was done using raw scores.

DIF Analysis

DIF analysis was carried out to identify possible group-specific items across the three tests. The current study used lordif version 0.3-3 (Choi et al., 2011), an R package, to conduct logistic ordinal regression DIF using IRT, which treats each item as ordinal dependent variables and uses IRT ability estimates (Expected a Posteriori) as the conditioning variable in ordinal regression analysis. DIF was examined with respect to gender (i.e., male vs. female) for all three tests. For the National test, the province (i.e., Hubei vs. Jiangsu) was also considered as a possible source of DIF. The presence of DIF items may not be directly related to establishing the equivalence of the three test forms, while results may inform readers as to how fair and valid each test form is to be used to make a high-stakes decision.

Results

Content Analysis

Test Specifications

In general, there was a close alignment among the three test specifications. They were similar in the following aspects: First, they all included a statement about the nature of the examination, a cognitive framework, an ability framework, a list of topics to be covered along with the cognitive requirement for each topic,

Table 2

Contents Covered in the Three Tests

Contents	National	Hunan	Jiangsu
Compulsory Contents			
(1) Sets	N1		J1
(2) Functions and elementary functions I	N3	H3, H8, H10	J10
(3) Elementary functions II (trigonometric functions) and identical transformation of trigonometric functions	N6, N8	H9	J5, J13, J15
(4) Solving triangles	N16	H18	J14, J18[a]
(5) Planar vectors	N15	H16	J12
(6) Number sequences	N17	H20	J7, J20
(7) Inequalities	N9	H14	
(8) Complex numbers	N2	H1	J2
(9) Derivation and its application	N11, N21	H22	J11, J19, J23
(10) Algorithms	N7	H6	J3
(11) Logic terms	N14	H5	
(12) Reasoning and proof			
(13) Probability and statistic I	N5	H2	J4, J6
(14) 3-D geometric shapes	N12	H7	J8
(15) Positional relationships between a point, a line, and a plane			J16
(16) Planar analytic geometry			J9
(17) Conic curves and their equations	N4, N10, N20	H15, H21	J17
(18) 3-D vectors and solid geometry	N19	H19	
(19) Counting principles	N13	H4	
(20) Probability and statistics II	N18[a]	H17[a]	J22
Elective contents			
(1) Selected lectures in geometric proof	N22	H12	J21(A)[b]
(2) Coordinate system and parametric equation	N23	H11	J21(C)[b]
(3) Selected lectures in inequalities	N24	H13	J21(D)[b]
(4) Matrices and transformations			J21(B)[b]

[a] These are application items. [b] There were four questions in J21, they were coded as J21(A), J21(B), and so forth. Students were required to select two of them to answer.

and so forth. Second, they all declared that the nature of the exam is a college admission exam for high school graduates and candidates with similar educational backgrounds. Third, the cognitive frameworks are the same in that they all have three levels: knowing, understanding, and mastering. Fourth, the skills to be evaluated include spatial imaginary skills, abstract thinking skills, reasoning/deductive skills, computational skills, data handling skills, and creative and application skills. Fifth, the content lists to be measured are similar for both compulsory and elective topics. The main differences among them are: First, compared with the national examination syllabus issued by NEEA, the Hunan version is an expanded version with exemplary questions and solutions added for each topic, whereas the Jiangsu version is a simplified version with the content areas and the cognitive requirements listed in a table. Second, the Hunan version has a separate section on mathematics ideas and methods.

Third, there is a section on the format, the structure of the test, and a practice test in the Hunan and Jiangsu versions. However, they are not included in the national examination syllabus issued by NEEA. Fourth, the Jiangsu version had an extra elective topic of Matrices and Transformation that covered second-order matrices and their applications to planar transformation.

Item Review

The examination of the test specifications indicated that these tests were designed to test mathematical abilities with items in similar content areas. The content area for each item was identified and listed in Table 2. For those blank cells in Table 2, the corresponding content areas were actually embedded in other content areas. For example, for the Hunan test, "Sets" was covered in H6 as part of the "Algorithm" item, which was identified as the main content area. It seems apparent that all the test developers tried to cover almost all content areas specified by the syllabi, although the number of items under each content area is slightly different. Due to the limited administration time (2–2.5 hours, 0.5 hours are for the supplementary section of the Jiangsu test), no more than 25 items could be used. Since most of the content areas need to be covered in one test, only 1–2 items were constructed for each content area.

As aforementioned, the Hunan test was selected for the study because it was similar to the National test for high school mathematics teachers. The data shown in Table 2 have also revealed this point. It was interesting to note that even for a unique application item, the related content area was the same, "Probability and Statistics II." For the Jiangsu test, the only application item is related to plane geometry and trigonometric functions.

Unidimensionality Assumption Check

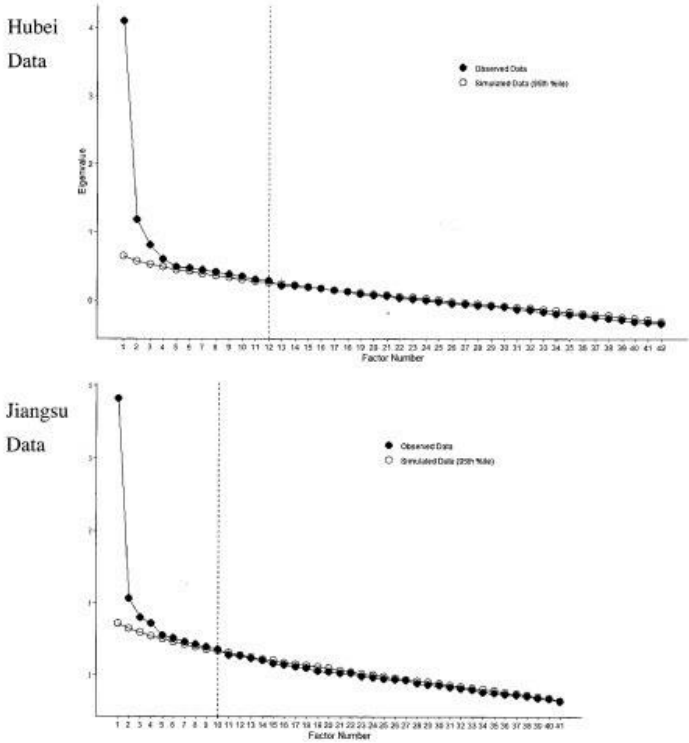
Prior to conducting any type of IRT analysis, it is important to ensure that the key assumptions of IRT are met by data. One of the essential assumptions of IRT is the unidimensionality assumption which requires that a test measures only a single latent trait. To assess the extent to which this assumption was tenable, a scree plot was visually inspected (Figure 1). To aid the interpretation of the finding, the parallel analysis was also performed using psych version 2.0.8 (Revelle, 2020), an R package, and the results are displayed in Figure 1. Based on the scree plots, both tests seemed to have a dominant factor, which explained 10.72% (Hunan) and 11.02% (Jiangsu) of the variance, while two or three additional factors suggested significance, creating the elbow point after the third (Hunan) or fourth (Jiangsu) factor. Note that the data used for this unidimensionality assessment were the merged dataset of the National test and either the Hunan or Jiangsu test, both of which were submitted for concurrent calibration simultaneously. Concurrent calibration operates under the assumption that there is a dominant single factor shared by two tests. According to the parallel analysis, 12 and 10 factors were diagnosed to be significant for Hunan and Jiangsu data, respectively, suggesting the potential violation of the unidimensionality assumption. However, the literature generally suggests the minimal impact of the unidimensionality assumption violation on parameter estimates, particularly when one dominant latent trait exists, and all other traits are nuisance factors (Drasgow & Parsons, 1983; Harrison, 1986; Junker & Stout, 1994), which was the case in the current study. Thus, we proceeded with the IRT analyses, and the results are presented below.

Item Analysis

The fact that the National test was administered to both groups (i.e., a single group design) enables the student abilities and item difficulties to be located on a single scale. Therefore, once the tests have been calibrated, the item parameter estimates from IRT analysis can be compared directly across the three tests. After this calibration, the Hubei group turned out to outperform the Jiangsu group and be more variable with its mean of .4 and the standard deviation of 1.22. The following

Figure 1

Scree Plots



discussion will be made based on estimates obtained from the concurrent calibration unless otherwise noted elsewhere.

Note that the use of 1PL resulted in a single a-parameter within a single form that is shared by all the items in the test. Therefore, each item had one common a-parameter plus its unique b-parameter. Item analysis results based on IRT are provided in Tables 3 and 4 for dichotomous and polytomous items, respectively.

Prior to examining item analysis results, the goodness-of-fit of the IRT model was evaluated using several statistics, including the root mean square error of approximation (RMSEA), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) based on limited information statistics, M_2. According to Maydeu-Olivares and Joe (2014), a cutoff value of .089 for RMSEA indicates an adequate fit and a value of .05 indicates a close fit. The RMSEA value for our data was found to be .02, suggesting a satisfactory model fit. Also, AIC values for the fitted model and zero-factor null model were shown to be 54,836.98 and 57,388.86, respectively. Also, a smaller RMSEA associated with the fitted model suggests an improved fit resulting from the use of the IRT model. Similarly, BIC values were 55,486.47 and 58,012.98 for the fitted model and zero-

Table 3

Item Analysis for Dichotomous Items

National			Hunan			Jiangsu		
Item	*a* (disc.)	*b* (diff.)	Item	*a* (disc.)	*b* (diff.)	Item	*a* (disc.)	*b* (diff.)
N1	.3952	−5.1288	H1	.4832	−2.7546	J1	.5533	−4.2351
N2		−4.0798	H2		−1.6554	J2		−4.3181
N3		−2.9747	H3		−3.1332	J3		−4.1576
N4		−1.5151	H4		−2.4961	J4		−4.4074
N5		−4.0798	H5		−3.0460	J5		−3.0115
N6		−1.6092	H6		−3.6608	J6		−4.0848
N7		−4.0989	H7		−2.1076	J7		−2.7851
N8		−2.8645	H8		−.1113	J8		−2.6595
N9		−2.1652	H9		−.0487	J9		−3.2088
N10		−2.0679	H10		1.0270	J10		−1.0487
N11		−.2212	H11		−1.1822	J11		−2.6595
N12		.7740	H12		−4.0722	J12		−.3295
N13		-	H13		−3.1714	J13		−.8538
N14		−5.8201	H14		−1.9154	J14		.4574
N15		−2.9622	H15		−.6248			
N16		−.6217	H16		2.4415			

Note. disc. indicates the item discrimination. diff. indicates the item difficulty.

Table 4

Item Analysis for Polytomous Items

Item	*a* (disc.)	*b*1	*b*2	*b*3	*b*4
N17	.3952	−2.7709	−2.5817	.6650	3.3321
N18		−.4282	.8302	2.0522	2.8293
N19		−3.1657	1.5158	2.3430	2.8353
N20		−1.7948	1.2964	2.6037	3.1236
N21		−.4282	.8302	2.0522	2.8293
N22		−6.1990	3.1008	3.5935	-
N23		−4.2506	−1.6375	4.6539	4.7420
N24		−.5805	−.1036	2.7768	3.5890
H17	.4832	−5.565	−3.577	−3.011	−1.451
H18		−4.614	−.5430	.3280	.7920
H19		−2.291	1.272	2.368	2.828
H20		−1.555	4.043	6.394	7.445
H21		−.247	4.174	6.414	7.108
H22		−.1110	2.652	5.417	7.918
J15	.5533	−5.8823	−4.1123	−2.2739	−1.2708
J16		−5.6049	−5.3640	−4.0858	-
J17		−3.7500	.4378	.9659	1.2848
J18		−1.1636	.7704	1.4146	1.9021
J19		−1.4614	.2230	3.7243	5.4531
J20		.0310	2.8217	5.4537	5.6984

Note. disc. Indicates the item discrimination. Each *b*-parameter estimates the minimum ability score to get a score of 1, 2, 3, and 4, respectively, with a probability of .5.

factor null model, respectively, supporting the good fit of the model to the data.

National Test

The estimate of the *a*-parameter for the National test is .3952. Lord (1975; as cited in Plake & Kane, 1991) reported a minimum of .25 and a maximum of approximately 2.5 for the *a*-parameter in a verbal section of the SAT. Therefore, the *a*-parameter for the National test did not seem satisfactory, particularly considering that the major purpose of the exam was to make the college admission decision, which usually requires a test to be highly discriminative.

A negative value of a *b*-parameter estimate suggests that the dichotomous items were relatively easy for the student population. It seemed apparent that most items were relatively easy for the students (the average value of b-parameters of −2.6290), except for N12, whose *b*-parameter estimate was .7740. Note that Table 3 reports item parameter estimates for dichotomous items only. Five items (N1, N2, N5, N7, and N14) were identified as extremely easy with their *b*-parameter estimates less than −4. Plake and Kane (1991) set boundaries of −2.0 and +2.0 for the *b*-parameter in their simulation study. Also, Nettles (1995) points out that the *b*-parameter typically ranges from −3.0 to +3.0. Consideration of the literature led to the conclusion that these five items in the National test were too easy.

The open-response items include multiple *b*-parameters, as noted in Table 4. Each *b*-parameter represents the difficulty level for each category. For example, N17 has four *b*-parameters of −2.7709, −2.5817, .6650, and 3.3321. Each estimate indicates the minimum ability score to get a score of 1, 2, 3, and 4, respectively, with a probability of .5. In order to earn a score of 1 on this item with a probability of .5, for instance, an ability score of −2.7709 is required. Of course, the higher the ability score, the more able the student is. Note that N22 has three *b*-parameter estimates, and it originally had a score range of 0–4. Some data manipulation was made for N22 because the occurrence of scores 0 and 1 was too rare to estimate the *b*-parameters separately, so these two score categories were combined and treated as a single category in the analysis.

Hunan Test

The *a*-parameter for the Hunan test was estimated as .4832, which is slightly higher than that for the National test. However, the discrimination parameter estimate still failed to reach an adequate level given that there needs to be adequate discrimination of scores at the higher end of the distribution when the purpose is to make an accurate decision regarding college admission. An examination of *b*-parameter estimates showed that the dichotomous items were generally easy with many negative estimates, except for H10 and H16, whose *b*-parameter estimates were 1.0270 and 2.4415, respectively (the average value of −1.6570). Three open-response items (H20, H21, and H22) seemed very difficult for students to receive a score of 2 or more, as the *b*-parameters for these categories were found to be high. Consequently, many students would end up with a score of 0 or 1 on them.

Jiangsu Test

The *a*-parameter for the Jiangsu test was found to be .5533, the highest estimate among the three tests. However, those three tests showed a similar level of item discrimination in general, which calls for a higher level of discrimination power in consideration of the primary purpose of NCAE. In terms of *b*-parameter estimates, more negative values were found for dichotomous items in the Jiangsu exam, with only one barely positive estimate for J14 (.4574). The average *b*-parameter value was found to be −2.6644. Among the polytomous items, J15 and J16 were relatively easy with all *b*-parameters being negative, and J19 and J20 were the most challenging items, especially in order to get a maximum score on them.

The *a*-parameters for the three tests were all below .6, which is low (Baker, 2001).

Considering the major purpose of NCAE, the discrimination level needs to be improved. Among the dichotomous items in the three tests, only four out of the 46 items had positive *b*-parameters. Seventeen of them (37%) even had *b*-parameters less than −3. Among the polytomous items, three of them (H17 and J15–16) were relatively easy with all *b*-parameters being negative, seven items (N17, N20, H20-22, J19–20) were challenging with *b*-parameters greater than 3 in order to get a full score on them. In summary, the three tests were similar in that they all involved many easy items but few challenging items with less discriminating power than desired. Therefore, in order to better serve the selection purpose of NCAE, adding more challenging and discriminating items is recommended.

Item Information Function (IIF)

The item information functions for each item are displayed in Figure 2 separately for three levels of item difficulty to avoid visual complexity. Criterion used to categorize items into "easy," "moderate," and "difficult" is an ability (θ) score below −1, between −1 and +1, and above +1.

National Test

Eleven items were identified as easy and were presented in the leftmost plot of the first row of Figure 2. Among them, 10 are dichotomous items, tending to have their peaks near the left tail of the ability score scale, indicating that they provide maximum information for examinees whose ability scores are low. In other words, these items are efficient in discriminating among people who have limited knowledge measured by the test. One remaining item (N23) is the open-response item that revealed two peaks near −3 and +4. This item generally provided more information than the others, as its IIF was above the other items. This is probably because it has multiple score categories collecting more information about examinees compared to MC items with a single score.

There were six items categorized as moderate items, and their IIFs can be found in the middle of the first row in Figure 2. They tended to show a similar pattern, providing maximum information near the ability score of 0. Six items classified as difficult were all open-response items, suggesting that the open-response items relatively worked better in discriminating among high-performing students than the dichotomous items. Additionally, the amount of information offered by these items tended to be similar over ability scores.

Hunan Test

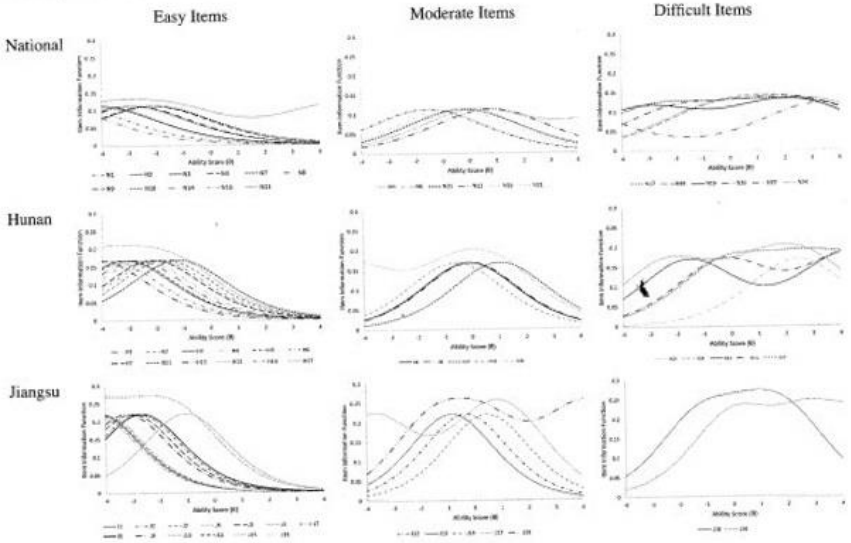
The IIFs for the Hunan items are provided in the second row of Figure 2. More than half (12) of the items were classified as easy items, providing maximum information on those with low ability scores. Particularly, a large amount of information was offered by one open-response item (H17) for those who had an ability score less than −2. Five items were assigned to the "moderate item" category. Besides the middle part of the scale, some additional information was promised with H18 for the lower part of the score scale. The remaining four items tended to maximize their information near 0. Five items were categorized as difficult items. Except for H16, the items considered as difficult were all open-response items. Unlike H22, which showed a monotonic increasing pattern with the ability score, H19 and H20 revealed two peaks, implying that these two items could discriminate well among both low- and high-performing test takers, but not middle-achieving examinees.

Jiangsu Test

The last row of Figure 2 contains three plots of the IIFs of the Jiangsu test. Thirteen items (J1–J11, J15, and J16) appeared as easy, having their IIF peaks below the ability score of −1. More information could be gained with J15, compared to J16 which did not show much difference with any other dichotomous items in terms of the amount of information, possibly due to its easiness. More specifically,

Figure 2

Item Information Function



when every examinee gets an extremely easy item correct, the addition of the item to a test does not help differentiate examinees with high ability from those with low ability. This becomes more problematic when an item is not a simple question, but an open-response item that requires more administration time. Even with the increased testing time, the item may not offer much information about the test takers, implying its inefficiency.

The IIFs for five items categorized as moderate items revealed that three dichotomous items (J12–14) had their peaks near 0, the other two polytomous items (J17 and J19) had two peaks, one near the middle of the scale, and one either at the bottom (J17) or at the top of the score scale (J19). Another two open-response items (J18 and J20) were assigned into the "difficult item" category, producing maximum information at the middle through upper part of the ability score.

The IIF curves of the items in the three tests indicated that they served best for students with low ability levels. The number of easy items in the National, Hunan, and Jiangsu tests accounted for 50%, 55%, and 65% of the total items, respectively. Among the 36 easy items, most of them had their peaks near the ability score less than −1, indicating that they provided maximum information for examinees with very low abilities.

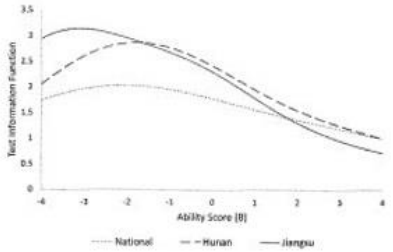
Three items at the easy, moderate, and difficult levels from each test are included as examples in the Appendix.

Test Information Function (TIF)

The magnitude of information provided by each test along the ability score was found by summation of IIFs of items in each exam (Figure 3). The examination of TIFs allowed for a direct comparison among the three exams. All three tests exhibited their peaks near the lower end of the scale, meaning that they were good at discriminating among under-performing

Figure 3

Test Information Function



students but less efficient for high-performing students. Particularly, the Jiangsu exam revealed a sharp decreasing pattern, having its maximum near the ability score of −3. The ability score was set to follow a standard normal distribution with a mean of 0 and a standard deviation of 1. Considering this property of the ability distribution, the Jiangsu test serves best for students who are at the bottom 0.1% of the entire examinee population. The Hunan exam showed its peak near the ability score of −2 and with the same logic of a standard normal distribution, this test best discriminates among students whose ability level are at the bottom 2.3% of the population. Among the three tests, the National test tended to offer the least information across the scale, and its fairly flat curve indicated that its discrimination power remained relatively constant across the score scale.

Reliability Analysis

Reliability estimates for the three tests are presented in the left part of Table 5. Notice that there are two distinct values for the National test for the α coefficient because reliability was estimated separately for each group: Hubei and Jiangsu students. As expected, reliability estimates based on the IRT framework showed slightly higher values compared to their counterparts of the α coefficient. This is because the IRT framework conceptualizes each form to be fixed, whereas Cronbach's α is computed under the assumption that forms (items) are essentially tau-equivalent (Graham, 2006). Due to the similar tendency between the coefficient and IRT-based reliability coefficient, the subsequent discussion will be made based on the α coefficient.

A higher reliability estimate was found for the Hubei students than for the Jiangsu students. In general, the three tests revealed a similar level of reliability ranging from .59 to .68, which seems relatively low given the high-stakes nature of the test (Spitzer et al., 1967). One potential cause of such low reliability was found with the limited variability in examinees' response data. As noted earlier, the sample of the current study was taken from two top-tier schools, and these students tended to be higher achievers in general. The homogeneity of the sample could potentially lead to low reliability of the tests. Considering that reliability is sample-dependent, further research is needed to examine the reliability of these tests with a broader student population. Another potential reason for low reliability is that if content areas are too distinct from one another, multiple abilities are required to earn a high score on the test. If this were true, then the stratified coefficient alpha would have yielded a more accurate and defensible value of reliability (probably higher reliability).

Correlation Analysis

Results for correlation analysis are shown in the rightmost columns of Table 5. The correlation between the Hunan and Jiangsu tests could not be observed because there was no examinee taking both forms. Prior to conducting the correlation analysis, raw scores used for the correlation analysis were examined with respect to the mean and standard deviation. The Hubei group showed the mean of 91.26 and 87.61 and the standard deviation of 27.37 and 18.61 for the National and Hunan tests, respectively. The two tests had the possible score range of 0–145 (after eliminating one dichotomous item) and 0–150. For the Jiangsu group, the means were 78.23 and 105.33, and the standard deviations were 16.27 and 18.09, with the possible score range

Table 5

Correlation and Reliability for Three Tests

	Reliability analysis		Correlation analysis		
	α coefficient	IRT reliability	National test	Hunan test	Jiangsu test
National test	.65 / .59	.64	1.00	.840	.943
Hunan test	.68	.74	.558	1.00	-
Jiangsu test	.64	.69	.579	-	1.00

Note. (a) The lower off-diagonal elements are Pearson correlations; the upper off-diagonal elements are disattenuated correlations; (b) For the National test, two values for the reliability estimate are obtained from Hubei students and Jiangsu students, respectively, for the α coefficient.

of 0–135 (due to the removal of supplementary section) and 0–160 for the National and Jiangsu tests, respectively.

In Table 5, the lower off-diagonal elements are the observed Pearson correlations between the tests, and the upper off-diagonal elements are disattenuated correlations that are corrected and estimated in a fashion that accounts for unreliability of measurement. For both the Hunan and Jiangsu tests, the correlation coefficients with the National test demonstrated a moderate level of association with r =.558 and .579, respectively. A significant relationship between the National and Jiangsu tests was noted with the disattenuated correlation of .943. Another strong, yet less correlated, relationship was found between the National and Hunan tests with the disattenuated correlation of .840. In sum, the three tests revealed moderate correlation values, partially supporting the equivalence among them.

Differential Item Functioning (DIF) Analysis

Results from the DIF analysis flagged two items of the National test (N6 and N20), two items of the Hunan test (H17 and H22) but no item of the Jiangsu test when tests were examined with respect to gender. For the National test, conditioning on the ability level, male students were likely to find N6 more difficult. N20 was found to be easier for male students whose ability level was high (to get a score of 3 or 4), while the reverse pattern was true for those whose ability level was low (to get a score of 1 or 2).

For the Hunan test, H17 was found to be easier for male students across all score categories, while H22 was found easier for male students to endorse the first category but more difficult to endorse the other score categories compared to their female counterparts. This suggests that given the same level of ability, male students were more likely to receive a higher score on H17 and to endorse a score of 1 on H22 while the reverse pattern was true for female students to receive a score of 2 or higher on H22.

Interestingly, when the province was used to detect DIF, half of the items of the National test were flagged, including items N3, N6, N9, N15–16, and N18–21. Among those, the following items were found to be easier for the Hubei students: N6, N15, and N18, whereas N3, N9, N16, and N19 were easier for the Jiangsu students. The remaining items produced a mixed result depending on the score category.

Impact of not Using a Single Group Design (Concurrent Calibration vs. Separate Calibration)

To examine the benefits of using the single group design for this type of research, separate calibration was carried out for each of the three test forms. Due to the IRT scale indeterminacy property, separate analyses led to an IRT scale with a mean of 0 and standard deviation of 1 for each group (despite the difference between groups), creating two IRT scales for each group that are not comparable. Specifically, results from concurrent calibration revealed that the student group taking the Hunan test had a

higher mean of ability ($\hat{\sigma}_\theta = 0.40$) compared to the Jiangsu student group ($\hat{\sigma}_\theta = 0.00$). Also, the former had a more variability ($\hat{\sigma}_\theta = 1.22$) than the latter ($\hat{\sigma}_\theta = 1.00$). These group differences could not be observed when separate calibration was used because each group was set to have the same mean and standard deviation for estimation purposes.

Another implication of using the single group design is that item parameter estimates can be compared across forms. Separate calibration resulted in estimates that are slightly different from those obtained under the single group design, particularly for the Hunan test. The a-parameter for this test was found as 0.5601 which was higher than 0.4832 found under the single group design. The value of 0.5601 should not be interpreted that this form had a higher a-parameter than the Jiangsu test (see Table 3; *a*-parameter of .5533) because these parameters are not on the same scale. Similarly, the *b*-parameters from separate calibrations ranged −3.8178 and 1.7759 while those ranged −4.0722 and 2.4415 under the single group design. This difference is not simply estimation errors but is primarily attributable to the fact that the IRT scale was specified differently. Thus, if one attempts to compare the Hunan test items with the Jiangsu test items, one should refer to parameter estimates obtained under the single group design. Without establishing a common IRT scale, a comparison between persons or items would not be valid.

Discussion

This study examined the equivalence of the three mathematics tests used in NCAE 2014 in China. The conclusion is two-fold: (a) the three tests are equivalent in terms of the content being primarily assessed, while (b) they were fairly distinct in several aspects, including item discrimination levels, the test information functions, and so forth. In terms of test content, their test specifications proved that the main contents and the core abilities being tested were comparable. Students from Hubei and Jiangsu provinces were taught the same content in mathematics so that they would be prepared for their higher education no matter which college they were admitted to. The conclusion suggests that the one-syllabus-multiple-tests practice seems to keep the same core educational mission and values while giving the local provinces their own autonomy.

However, the three tests were different in various aspects. First, the three tests revealed slightly different levels of item discrimination. The Jiangsu test demonstrated the highest level of item discrimination in general, followed by the Hunan and National tests. The highest item discrimination value of the Jiangsu test was probably due to its format of dichotomous items. They were all set as short-answer questions, which did not allow the participants to use a blind guessing strategy as they might have done in answering MC questions.

In addition, the TIF curves for the three exams were different appreciably. The TIF curve for the National test was lower and relatively flatter than those for the Hunan and Jiangsu tests, meaning that it offered the least information constantly across the ability scale. A similar trend was observed in the IIF curves with peaks for the items in the National test lower than those in the Hunan and Jiangsu tests. Except for these differences, all three tests served best for students who are at the bottom of the entire population with respect to mathematics ability.

Both the discrimination and the reliability values for the three tests were low. The item review revealed that although the three tests consisted of items similar in content areas, most of the time, the number of items in one content area in one test is only one. They might be different for the test-takers though they are often taken as "mathematical" problems (Carroll, 1996). The study by Carroll (1996) found that mathematical tasks differ over the various branches of mathematics and even within the same branch like arithmetic. Results from this study also suggest that there were

many items in the mathematics tests with low difficulty levels as well as low discrimination values. Although the inclusion of easy items may relieve test anxiety, having too many easy items does not appropriately serve the purpose of such high-stakes examinations to select candidates for higher education. An item to simplify a fractional expression with i was often included in NCAE mathematics tests since 1980 (Swetz & Chi, 1983). As the interpretation and fair use of test scores are the most important components of validity (Messick, 1995), more open-response items were suggested to provide more information about the test-takers. The 2017 version of *Mathematics Curriculum Standards for High Schools* (MOEPRC, 2017) also suggested reducing the number of MC and short-answer questions in Gaokao, and increasing the number of open-response items that measure students' thinking skills. Since the national test will be used in more provinces in the coming years, its items need to be improved to provide more information. Recall that the Jiangsu test had a supplementary section, particularly for students in the science stream, and the national mathematics test may take a similar route to deal with those difficult topics to meet the needs of students who pursue higher degrees in science, technology, engineering, and mathematics (STEM) fields.

Also, the local tests seemed to be robust to differential item functioning with none or two DIF-flagged items, while a couple of items were detected to be potentially biased for the National test with respect to the gender and province variables. A thorough investigation of items would be encouraged to equally treat students from all different demographic backgrounds.

Finally, this paper contributed to the literature by shedding light on the benefits of using a single group design. Specifically, separate calibration (without the single group design) produced slightly different results from those under the single group design and did not allow us to make a fair comparison between the two test forms. This issue could be overcome successfully by employing the single group design.

While the results of the current study showed some interesting and promising findings in regard to the equivalence of the three tests, limitations are worthy of note. The students were sampled from so-called "top-tier" schools, which probably lowered the b-parameters estimated for those students, while we assumed that the sample could be regarded as similar to the target student population (the entire students taking the test) as our sample took the tests one year prior to their actual testing. Therefore, further study is needed to evaluate if the same results are replicable using data from a different province or a different school setting. Rather, the current study primarily focused on illustrating how to evaluate national and local-level tests using IRT under the single group design while providing some empirical evidence of test equivalence for selected tests.

Conclusion

The conclusion of the current study is that the three tests used in NCAE 2014 in China are equivalent, and they were also different in a variety of aspects. The study by Jiang et al. (2019) indicated that the National test was not sensitive enough to distinguish between low- and high-performing students. The current study provides insightful information on how to improve the validity of National mathematics test scores by comparing them with scores from the Hunan and Jiangsu tests. Hunan and Jiangsu provinces started to use the National test again in 2016 and 2021, respectively. As mentioned earlier, multiple-choice items with multiple responses and ill-structured items were included in the national test since 2020, whether the inclusion of the new categories of items could potentially improve its capability of providing more information for students at moderate and high ability levels needs to be further studied.

Nowadays, more high school graduates are trying to pursue their undergraduate degrees away from their home countries. Similar studies that examine the equivalence of multiple test

forms can be conducted across countries using a single group design. We can first examine whether their high school mathematics curriculum standards (and/or examination syllabi) are similar, particularly whether the topics covered are comparable. Then, we can collect students' data in various countries with students from a specific country tested by their local test and a "common" test that might be in English and easy to translate. Finally, we can do the data analysis in a similar way as the current study. This information is important for higher education institutions for the design of their undergraduate programs.

References

Ai, H., & Zhou, Y. (2017). 基于 SOLO 分类理论的高考数学试题思维层次分析：以 2016 年全国卷（理科）为例 [Questions of mathematics based on SOLO taxonomy theory: Taking the national science volume in 2016 as an example]. *Educational Measurement and Evaluation*, *2017*(5), 58–64.

Baker, F. B. (2001). *The basics of item response theory* (ED458219, 2nd ed.). ERIC. https://files.eric.ed.gov/fulltext/ED458219.pdf

Bao, J. (2002). 中英兩國初中數學期望課程綜合難度的比較 [A comparative study on the comprehensive difficult of Chinese and British intended mathematics curricula]. *Global Education*, *31*(9), 48–52.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press. https://doi.org/10.1016/C2013-0-10375-3

Cai, L. (2018). *flexMIRT* (Version 3.5) [Computer software]. Vector Psychometric Group. https://vpgcentral.com/software/flexmirt/

Carroll, J. B. (1996). Mathematical abilities: Some results from factor analysis. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 3–25). Routledge. https://doi.org/10.4324/9780203053270

Chen, X., Zhou, J., Shi, D., & Pan, Y. (2022). 基于 SOLO 分类理论的高考数学试题研究 [Research on national college entrance examination maths tests based on SOLO classification theory]. *Journal of Baicheng Normal University*, *36*(2), 111–119.

China Education Online. (2021, June 7). *2021 高招报告：我国高等教育规模达到世界第一，高招录取率突破 90%* [2021 Report on college admission: The scale of China's higher education ranked 1st and the admission rate has reached 90%]. EOL. https://news.eol.cn/yaowen/202106/t20210607_2119442.shtml

China Education Yearbook Editorial Board. (2014). *中国教育统计年鉴* [China education yearbook 2013]. People's Education Press.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*, 1–30. https://doi.org/10.18637/jss.v039.i08

Chu, R., Wang, J., Wang, Z., & Ding, Y. (2005). 2005 年高考数学试卷分析 [An analysis of the mathematics examination papers for Gaokao 2005]. *China Examinations*, *2005*(11), 23–38.

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. https://doi.org/10.1080/03098770701625761

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, *73*(2), 24–32.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional date. *Applied Psychological Measurement*, *7*(2), 189–199.

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. https://doi.org/10.1080/10670569508724213

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944. http://dx.doi.org/ 10.1177/0013164406288165

Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory Into Practice*, *42*(1), 66–74. https://doi.org/10.1207/s15430421tip4201_9

Gu, M., Ma, J., & Teng, J. (2017). *Portraits of Chinese schools*. Springer Singapore. https://doi.org/10.1007/978-981-10-4011-5

Han, J., Yang, Z., & Wang, P. (2022). 中美高考数学试卷比较研究 [Comparative study of mathematics test papers in college entrance examination between China and the United States]. *Journal of Mathematics Education*, *31*(2), 13–20.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*(2), 91–115. https://doi.org/10.2307/1164972

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. https://doi.org/10.3868/s060-003-014-0010-7

Hunan Education Examination Authority. (2014). *2014 年湖南高考理科数学考试说明* [Description of Hunan Version of the 2014 National Higher Education Entrance Examination].

Jiang, C., Kim, D.-H., Wang, C., & Wang, J. (2019). Premises and challenges of high-stakes examinations: National higher education entrance examination mathematics test scores in China. *Journal of Applied Educational and Policy Research*, *4*(1), 1–21.

Jiangsu Education Examination Authority. (2013). *2014 年江苏省高考说明* [Description of Jiangsu version of the 2014 National Higher Education Entrance Examination]. Jiangsu Education Press.

Junker, B. W., & Stout, F. W, (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 51–84). Edumetrics Research Group, University of Ottawa.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

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). Information Age Publishing.

Li, B., & Shi, Y. (2020). 中国大陆和台湾地区高考数学试题难度比较研究：以2016–2018 年大陆全国卷 I 与台湾指考试题为例 [Comparative study on the difficulty of mathematics test in the college entrance examination between Chinese Mainland and Taiwan: Take the 2016 to 2018 Mainland national examination paper I and Taiwan development required test paper as a case]. *Journal of Mathematics Education*, *29*(1), 58–64.

Lim, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102. https://doi.org/10.1207/s15324818ame0601_5

Liu, H. (2015, March 12). 高考命题从分到统的历史逻辑 [Historical logic of the development of Gaokao from individualization to unification]. *中国教育报* [China Education Daily]. http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/

jyzt_2015nztzl/lianghui/pinglun/202103/t20210329_523321.html

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 programs*. Routledge. https://doi.org/10.4324/9780203056615

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

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. https://doi.org/10.1037/0003-066X.50.9.741

Ministry of Education of People's Republic of China. (2003). 普通高中数学课程标准（实验）[Mathematics Curriculum Standards for High Schools (Trial version)]. *People's Education Press*.

Ministry of Education of People's Republic of China. (2017). 普通高中数学课程标准（2017 年 版）[Mathematics curriculum standards for High Schools (2017 version)]. *People's Education Press*.

Ministry of Education of People's Republic of China. (2019, September 30). *新高考過渡時期數學學科考試範圍說明* [Specification of the examination scope of mathematics of the new college entrance examination in the transitional period]. Sina. https://k.sina.com.cn/article_6588938277_188bb382501900iacq.html?from=edu

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. ETS Policy Information Center.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Boston College; TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2015/international-results/

National Education Examination Authority. (Ed., 2014). 2014 年普通高等学校招生全国统一考试大纲（理科数学）[Examination Syllabus for the 2014 National Higher Education Admission Examination (Mathematics for the science stream)]. *Higher Education Press*.

National Education Examinations Authority. (2016). 突出實踐性和創新性，實現高考的選拔功能：2016 年高考數學試題評析 [Carrying out the selection mission of the College Entrance Examination while emphasizing practicality and innovation: An evaluation of the math papers of the 2016 College Entrance Examination]. *Journal of China Examinations*, *291*, 12–15, 57.

National Education Examinations Authority. (2019a). *2019 年普通高等學校招生全國統一考試大綱* [Examination syllabi for the 2019 National College Entrance Examination]. https://www.neea.edu.cn/res/Home/1901/d722242b1b7b3b4eed7d217dc782789a.pdf

National Education Examinations Authority. (2019b). 中国高考评价体系 [Assessment system of China' Gaokao]. People's Education Press.

National Education Examinations Authority. (2020). 以評價體系引領內容改革 以科學情境考查關鍵能力 —2020 年高考數學全國卷試題評析 [Carry out content reform under Gaokao assessment framework and evaluate key competences based on proper contexts—Analysis of the mathematics test of 2020 Gaokao]. *Journal of China Examinations*, *340*, 29–34.

National Education Examinations Authority. (2021). 聚焦核心素養考查關鍵能力 —2021 年高考數學全國卷試題評析 [Focusing on core competences and assessing key abilities—Analysis of the mathematics test of the 2021 Gaokao]. *Journal of China Examinations*, *351*, 70–76.

National Education Examinations Authority. (2022). 創設情境發揮育人作用深化基礎考查核心素養 —2022 年高考數學全國卷試題評析 [Create contexts to strengthen moral education and deepen basic knowledge to evaluate core literacy: Analysis of the national mathematics test of the 2022 Gaokao]. *Journal of China Examinations*, *363*, 14–19.

Nettles, S. S. (1995). Future psychometric practices in licensure testing. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 321–345). Buros Institute of Mental Measurements.

Nguyen, T. N. Q. (2019). Vietnamese standardized test of English proficiency: A panorama. In L. I. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 71–100). Routledge.

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

Organization for Economic Cooperation and Development. (2016). *PISA 2015 Results (Volume I): Excellence and equity in education*. PISA; OECD Publishing. https://doi.org/10.1787/9789264266490-en

People's Daily Online. (2019, June 28). *Gaokao scores accepted by more overseas universities*. People's Daily Online. http://en.people.cn/n3/2019/0628/c90000-9592417.html

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement*, *28*(3), 249–256. https://doi.org/10.1111/j.1745-3984.1991.tb00357.x

Reshetar, R., & Pitts, M. (2020). General academic and subject-based examinations used in undergraduate higher education admissions. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admissions practices: An international perspective* (pp. 237–255). Cambridge University Press. https://doi.org/10.1017/9781108559607

Revelle, W. (2020). *psych: Procedures for personality and psychological research* (R package; Version 2.0.8)[Computer Software]. Northwestern University. https://CRAN.r-project.org/package=psych

Riddle, W. C. (2005). *Educational testing: Implementation of ESEA Title I-A requirements under the "No Child Left Behind Act"* (RL31407). Congressional Research Service.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(Suppl 1), 1–97. https://doi.org/10.1007/BF03372160

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. http://www.barrons.com/articles/u-s-colleges-put-chinas-gaokao-to-the-test-1446188818

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry*, *17*(1), 83–87. https://doi.org/10.1001/archpsyc.1967.01730250085012

Swetz, F., & Chi, A. Y. (1983). Mathematics entrance examinations in Chinese institutions of higher education. *Educational Studies in Mathematics*, *14*(1), 39–54. https://www.jstor.org/stable/3482305

Tan, C. (2017). Private supplementary tutoring and parentocracy in Singapore. *Interchange*, *48*, 315–329. https://doi.org/10.1007/s10780-017-9303-4

The State Council. (2014, September 3). *国务院关于深化考试招生制度改革的实施意见（国 发 [2014]35 号）*[Implementation Recommendations for Deepening the Reform of Examination and Admission System]. The State Council, the People's Republic

of China. http://www.gov.cn/zhengce/content/2014-09/04/content_9065.htm

Wang, H. (2010). Research on the influence of college entrance examination policies on the fairness of higher education admissions opportunities in China. *Chinese Education & Society*, *43*(6), 15–35. https://doi.org/10.2753/CED1061-1932430601

Wang, L. (2008). Rasch 测量原理及在高考命题評價中的實證研究 [Rasch measurement principles and implementation in the entrance examination to higher education evaluation]. *Journal of China Examinations*, *188*, 32–39.

Wang, X. B. (2006). *An introduction to the system and culture of the college entrance examination of China* (ED562597). The College Board. https://files.eric.ed.gov/fulltext/ED562597.pdf

Wang, Y., & Zhou, Y. (2020). 新课标背景下高考数学试题 SOLO 思维层次研究 — 以 2019 年高考数学全国卷为例 [Research on SOLO thinking level of national matriculation mathematics test under the background of new curriculum standard—Taking 2019 mathematics national volume for example]. *Educational Measurement and Evaluation*, *2020*(4), 17–24.

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

Wu, G. (2007). *高考效度问题研究* [Study on the issue of validity of National College Entrance Examination; Doctoral dissertation, Xiamen University].

Wu, X., & Zhang, Y. (2018). 中国和韩国高考数学试题综合难度比较研究 [Comparative study on the comprehensive difficulty of mathematics questions of college entrance examination in China and Korea]. *Journal of Mathematics Education*, *27*(3), 19–24.

Yang, G. (2021). A sudden termination of a thousand-year history: Scholar-officials accomplished by the imperial examination system and the system overturned by scholar-officials' criticisms. *Journal of East China Normal University (Humanities and Social Sciences)*, *53*(3), 67–95, 179–180.

Yang, X. (2007). 中国高考史述论 (1949-1999)[Historical review of national higher education entrance examination in China (1949-1999)]. *Hubei People's Press*.

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

Yu, T. (2022). 基于 SOLO 分类理论的高考数学多选题评价研究 [Analysis of multiple-choice multiple-responses items in Gaokao based on SOLO taxonomy theory]. *Correspondence of the Teaching of Mathematics*, *2022*(6), 6–8.

Zhang, H. (2015, June 26). *Italy, France accept gaokao scores*. China Daily. 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, Y., Wu, X., & Peng, N. (2016). 综合难度系数模型在 2016 年高考数学试题评价中的应用 [Applying the comprehensive difficulty model in the evaluation of mathematics items in NCAE 2016]. *Educational Measurement and Evaluation*, *2016*(12), 47–53.

Zhang, Y., & Zhou, X. (2020). 综合难度视角下中法高考数学试题的比较研究：基于 2015–2019 年中国和法国高考数学试 卷 [A comparative study of Chinese and French college entrance examination mathematics items (2015–2019): A focus on overall difficulty]. *Journal of Mathematics Education*, *29*(3), 43–50.

Zhao, W. (2020). Predicament and outlook of China's math education. *National Science Review*, *7*(9), 1513–1517. https://doi.org/10.1093/nsr/nwaa070

Appendix A

Three Items at Easy, Moderate, and Difficult Levels From Each Test

In this appendix, three items at easy, moderate, and difficult levels from each of the National test, the Hunan test, and the Jiangsu test were presented as examples.

Three Items From the National Test

N5. Four students can participate in a public service activity on either Saturday or Sunday, then the probability that the activity can be served on both days is

(A)$\frac{1}{8}$ (B)$\frac{3}{8}$ (C)$\frac{5}{8}$ (D)$\frac{7}{8}$

N16. Given that a, b, and c are the opposite sides of ΔABC internal angles A, B, and C, respectively, $a = 2$ and $(2+b)(\sin A - \sin B) = (c - b)\sin C$, the maximum value of the area of ΔABC is _____.

N20. Given that $A(0, -2)$, an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1(a > b > 0)$ with an eccentricity of $\frac{\sqrt{3}}{2}$, F is a focus of the ellipse, the slope of line AF is $\frac{2\sqrt{3}}{3}$, O is the origin.

(a) Find the equation of the ellipse.

(b) A moving line l through point A intersects the ellipse at points P and Q, find the equation of line l when the area of ΔOPQ reaches its maximum value.

Three Items From Hunan Test

H3. Given that $f(x)$ and $g(x)$ are even and odd functions with the domain $(-\infty, +\infty)$, respectively, and $f(x) - g(x) = x^3 + x^2 + 1$, then $f(1) + g(1) =$

(A) -3 (B) -1 (C) 1 (D) 3

H18. As shown in Figure A.1, in a planar quadrilateral $ABCD$, $AD = 1, CD = 2, AC = \sqrt{7}$.

(a) Find the value of $\cos\angle CAD$.

(b) If $\cos\angle BAD = -\frac{\sqrt{7}}{14}$, $\sin\angle CBA = -\frac{\sqrt{21}}{6}$, find the length of BC.

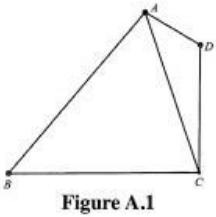


Figure A.1

H22. Given that a is a constant number that is larger than 0, and a function $f(x) = \ln(1 + ax) - \frac{2x}{x+2}$.

(a) Discuss the monotonicity of $f(x)$ on the interval $(0, +\infty)$;

(b) If function $f(x)$ has the extrema x_1 and x_2, and $f(x_1) + f(x_2) > 0$, find the range of values of a.

Three Items From Jiangsu Test

J5. Given two functions $y = \cos x$ and $y = \sin(2x + \varphi)$ $(0 \le \varphi < \pi)$, and their graphs intersect at a point whose x-coordinate is $\frac{\pi}{3}$, then the value of φ is _____.

J17. As shown in Figure A.2, in the Cartesian plane xOy, Points F_1 and F_2 are the left and right foci of an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1\,(a > b > 0)$, respectively. The coordinates of co-vertex B are $(0, b)$. Draw a segment BF_2 and extend it to intersect the ellipse at Point A. A line segment through point A is perpendicular to the x-axis and intersects the ellipse at Point C. Draw the segment F_1C.

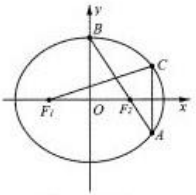


Figure A.2

(a) If the coordinates of Point C are $\left(\frac{4}{3}, \frac{1}{3}\right)$ and $BF_2 = \sqrt{2}$, find the equation of the ellipse.

(b) If $F_1C \perp AB$, find the eccentricity e of the ellipse.

J20. Let S_n be the sum of the first n terms of a number sequence $\{a_n\}$. For any positive integer n, if there always exists a positive integer m such that $S_n = a_m$, the sequence $\{a_n\}$ can be named "H sequence".

(a) If $S_n = 2^n\,(n \in N^*)$, show that $\{a_n\}$ is an "H sequence".

(b) If $\{a_n\}$ is an arithmetic sequence with $a_1 = 1$ and the common difference $d < 0$, find the value of d so that $\{a_n\}$ becomes an "H sequence".

(c) Prove: for any arithmetic sequence $\{a_n\}$, there always exist two "H sequences" $\{b_n\}$ and $\{c_n\}$ such that $a_n = b_n + c_n\,(n \in N^*)$.

Appendix B

flexMIRT syntex

Concurrent Calibration

```
<Project>
 Title="Concurrent Calibration_1PL";
 Description="Concurrent Calibration_1PL";

<Options>
 Mode = Calibration;
 SavePRM = Yes;
 SaveInf = Yes;
 SaveICC = Yes;
 FitNullModel=Yes;
 M2 =Full;
 MaxE = 6000;
 MaxM = 6000;
 FisherInf = 81 , 4.0;

<Groups>
%GH%
 File ="DataH.csv";
 Varnames = N1-N12, N14-N24, H1-H22;
 Ncats(N1-N12, N14-N16) = 2;
 Model(N1-N12, N14-N16) = Graded(2);
 Ncats(N17-N21) = 5;
 Model(N17-N21) = Graded(5);
 Ncats(N22) = 4;
 Model(N22) = Graded(4);
 Ncats(N23-N24) = 5;
 Model(N23-N24) = Graded(5);
 Ncats(H1-H16) = 2;
 Model(H1-H16) = Graded(2);
 Ncats(H17-H22) = 5;
 Model(H17-H22) = Graded(5);

%GJ%
 File ="DataJ.csv";
 Varnames = N1-N12, N14-N21, J1-J20;
 Ncats(N1-N12, N14-N16) = 2;
 Model(N1-N12, N14-N16) = Graded(2);
 Ncats(N17-N21) = 5;
 Model(N17-N21) = Graded(5);
 Ncats(J1-J14) = 2;
 Model(J1-J14) = Graded(2);
 Ncats(J15, J17-J20) = 5;
 Model(J15, J17-J20) = Graded(5);
 Ncats(J16) = 4;
 Model(J16) = Graded(4);

<Constraints>
 Free GH, Mean(1);
 Free GH, Cov(1, 1);

 Prior GH, (N1-N12, N14-N24), Slope: logNormal(0, 0.5);
 Prior GH, (H1-H22), Slope: logNormal(0, 0.5);
 Prior GJ, (J1-J20), Slope: logNormal(0, 0.5);

 Equal GH, (H1-H22), Slope;
```

```
Equal GJ, (J1-J20), Slope;

Equal GH, (N1), Slope : GJ, (N1), Slope:
GH, (N2), Slope : GJ, (N2), Slope:
GH, (N3), Slope : GJ, (N3), Slope:
GH, (N4), Slope : GJ, (N4), Slope:
GH, (N5), Slope : GJ, (N5), Slope:
GH, (N6), Slope : GJ, (N6), Slope:
GH, (N7), Slope : GJ, (N7), Slope:
GH, (N8), Slope : GJ, (N8), Slope:
GH, (N9), Slope : GJ, (N9), Slope:
GH, (N10), Slope : GJ, (N10), Slope:
GH, (N11), Slope : GJ, (N11), Slope:
GH, (N12), Slope : GJ, (N12), Slope:
GH, (N14), Slope : GJ, (N14), Slope:
GH, (N15), Slope : GJ, (N15), Slope:
GH, (N16), Slope : GJ, (N16), Slope:
GH, (N17), Slope : GJ, (N17), Slope:
GH, (N18), Slope : GJ, (N18), Slope:
GH, (N19), Slope : GJ, (N19), Slope:
GH, (N20), Slope : GJ, (N20), Slope:
GH, (N21), Slope : GJ, (N21), Slope:
GH, (N22), Slope : GH, (N23), Slope:
GH, (N24), Slope;

Equal GH, (N1-N12,N14-N21), Intercept : GJ, (N1-N12,N14-N21), Intercept;
Equal GH, (N17-N21), Intercept(2) : GJ, (N17-N21), Intercept(2);
Equal GH, (N17-N21), Intercept(3) : GJ, (N17-N21), Intercept(3);
Equal GH, (N17-N21), Intercept(4) : GJ, (N17-N21), Intercept(4);
```

Separate Calibration (National Test)

```
<Project>
 Title="separate Calibration_1PL";
 Description="separate Calibration_1PL";

<Options>
 Mode = Calibration;
 SavePRM = Yes;
 SaveInf = Yes;
 SaveICC = Yes;
 FitNullModel=Yes;
 M2 =Full;
 MaxE = 6000;
 MaxM = 6000;
 FisherInf = 81 , 4.0;

<Groups>
%GH%
 File ="DataN.csv";
 Varnames = N1-N12, N14-N21;
 Ncats(N1-N12, N14-N16) = 2;
 Model(N1-N12, N14-N16) = Graded(2);
 Ncats(N17-N21) = 5;
 Model(N17-N21) = Graded(5);

<Constraints>
 Prior GH, (N1-N12, N14-N21), Slope: logNormal(0, 0.5);
 Equal GH, (N1-N12, N14-N21), Slope;
```