International comparative studies suggest that students from countries with high-stakes examinations often perform better than students from other countries. Nowadays more and more countries, including the United States, are implementing high-stakes examinations at the national or state levels. In this paper, we use the National Higher Education Entrance Examination (NHEEE) in China as an example to illustrate the premises and challenges of high-stakes examinations. NHEEE, commonly known as *Gaokao*, is the only measure used in China to determine if and which college a high school graduate is admitted to. This study examines the reliability and validity of scores obtained from the 2002 mathematics test of this critical examination that determines the future of thousands of students in China. Results of the Rasch analysis indicated that the uni-dimensionality assumption was tenable. The results also showed that the item reliability and separation were satisfactory, but the person reliability and separation were low. The low person separation reliability indicates that the exam is not sensitive enough to distinguish between low- and high-performing students. Examination of the person-item map suggested a need for more items at the intermediate and difficult levels to improve the reliability and validity of the test scores and to match the students’ ability levels. Results showed that the majority of items displayed little or no DIF between male and female students. Predictive aspects of validity are also reported.

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

High-stakes examinations have been criticized for their inappropriate use of results (Wu & Hornsby, 2002), for their undesirable “backwash” or “trickle-down” on classwork and study of students at lower grades, and for their negative effect on students’ personality characteristic (Madaus, 1991). However, it is still widely used in many countries including China, India, Singapore, South Korea, Japan, etc. National Higher Education Entrance Examination (NHEEE), commonly known as Gaokao, is important in China for high school graduates because the results could determine not only what university they can enter, but also their future career (Davey, Lian, & Higgins, 2002; Lambert, 2002; Wang, 2002). NHEEE is the only measure used by university admission offices in China to evaluate their applicants. The scores of NHEEE determine whether a student can be admitted to a prestigious university or not and whether a student can study in the major he/she chooses. The job opportunities for college graduates are not the same for all college majors or institutions. Quite a large number of high school graduates could not have the opportunity to pursue higher education because of their low scores in NHEEE. In 2002, about 9,400,000 students took NHEEE on June 7-9 (China News, 2002), while only about 6-7% of them (approximately 600,000 students) could be admitted to the so-called prestigious universities (Xiong, 2002). It is not exaggerating to claim that NHEEE scores play a large determining factor in a student’s career choice, so families in China invest a lot of time and money into their children’s preparation of NHEEE.

The pressure even moves downward to elementary schools. Many families stop their fifth-grade children’s after-school activities, such as music and sports, so that their children can have time to attend after-school and/or weekend courses in the major academic areas, such as Chinese, mathematics, and English, hoping that their children could get admitted to prestigious middle schools. To these
families, prestigious middle schools prepare their children for top-tier high schools, in which the students are more likely to be admitted to prestigious universities (Yang, 2002). As students move up each grade level, pressure increases because they are getting closer to taking NHEEE. Despite the fact that NHEEE plays a central role in students’ education and future career success in China and that the interpretation and fair use of test scores are the most important component of validity (Messick, 1995), few studies have investigated the validity of these scores. The present study aims to fill this gap.

Validity Evidences of NHEEE

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

The theoretical framework that guided this study is Messick’s (1995) framework of validity. Validity is the degree to which a test measures what is supposed to measure, so the interpretation and use of the scores (consequential validity) is of utmost importance. The classical models of test validity include content validity, criterion-related validity, construct validity, consequential validity, and that criterion-related validity has two forms: concurrent validity and predictive validity.
Gender Differences in Mathematics Performance

Mathematics is often taken as a male domain (Hyde, Fennema, Ryan, Frost, & Hopp, 1990). As the gender gap in mathematics performance narrowed in recent years (Hyde & Mertz, 2002; OECD, 2002), Andreescu, Gallian, Kane, and Mertz (2002) argued that female students with high ability in mathematics can be identified and nurtured. However, comparative studies across countries suggest that male students outperform female students in more countries (OECD, 2002). Similar trends can be found in China. In 1957, only about 23% of the students in higher education institutions were female. From 1995 to 2002, the proportion of female students increased from 35.4% to 45.7%. The year of 2002 is the first year with more female students than male students admitted to higher education institutions (Wuhan Wanbao, 2002). This trend kept going in recent years with the difference in proportion of female and male students approaching 10% in 2002. It is interesting to note among the 63 students who received the highest scores in NHEEE in 27 provinces in 2002, 33 (52.4%) were female (Wuhan Wanbao, 2002). The proportion of female students whose scores were the highest in the respective provinces increased from 29% to 53% from 1952-1999 to 2000-2002 (Airuishen China’s University Alumni Association, 2002). More female students than male students were studying in the Department of Mathematics in both top-tier universities and normal universities in China (Yang, He, & Ning, 2002). In a provincial normal university, the ratio of male and female students is 1:3.67 (Song & Zhang, 2002). Female university students also performed better than male students in mathematics (Qiu, Chen, & Xiao, 2002). So how did male and female students compare in high school mathematics? The results from previous studies were inconsistent. Tian and Zhu (2002) used the NHEEE 2002-2002 data from Ningxia province and found that male students performed better than female students in mathematics. However, Ye (2002) used the NHEEE 2002-2002 data from Zhejiang province and found that female students performed
better than male students in mathematics in 2002-2002, but worse than male students in 2002. Wan (2002) found that male students performed better than female students in a high school in Sichuan. In terms of specific mathematics content areas, gender differences also existed. Ye (2002) found that male students performed better than female students in sets and simple logic, plane vectors, permutation and combination, but female students performed better than male students in trigonometry, conics, limits and differentiation, complex numbers, and analytic geometry. The inconsistency in mathematics performance between male and female students indicate that further evidence is needed.

**Data Analysis**

The 19 common items and three student self-selected items were concurrently calibrated using the dichotomous Rasch model (Rasch, 1960) and Masters’ Partial Credit Rasch model (PCM; Masters, 1982) in Winsteps software (Linacre, 2002). Estimating parameters for the common and student self-selected items simultaneously in a single calibration run assures that all parameter estimates are on the common scale. The student self-selected items that are not taken by a group of students are treated as not reached or missing (Lord, 1980). One item (item 13) was excluded from the current analysis because it was not correctly edited in the file that we downloaded from the internet. It is the first short-answer question, which is normally easier than item 14. The unidimensionality of the measure was examined using Rasch Principal Components Analysis of Residuals (PCAR) and item Mean Square (MNSQ) fit values as implemented in Winsteps. The MNSQ fit values between 0.6 and 1.4 were considered reasonable (Bond & Fox, 2002). A variance of greater than 50% explained by the Rasch dimension with the additional dimension accounting for less than 5% of the unexplained variance is considered adequate (Linacre, 2002). For polytomous items, the effectiveness of the rating scale category was evaluated based on the criteria outlined in Wolfe and Smith (2002). Reliability was examined using person and item reliability and person and item separation indices. The hierarchy of item difficulties and its relationship to person abilities were examined using the person-item map. Differential item functioning (DIF) analysis was performed between male and female students. The difference in item difficulty estimates (i.e., DIF contrast) greater than 0.5 logits with \( p < .05 \) is considered substantial (Linacre, 2002), and further investigation is warranted.
Pearson correlation coefficients and analysis of variance (ANOVA) were used to examine the predictive validity of the NHEEE scores. The 2002 NHEEE mathematics test scores were correlated with their 2002 NHEEE mathematics test scores. The 2002 ANOVA was used to see if significant differences exist in their performance on the NHEEE mathematics test in 2002 with respect to the tiers of universities they were admitted based on their performance in 2002 NHEEE total scores.

RESULTS

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

Of the three student self-selected items, item 23 seems to be the easiest item and the most popular choice. Item 23 is related to ellipse and line in analytic geometry, which was just learned in the past year. It is not surprising that a higher percentage of participants selected this item. Item 22 is related to geometrical proof, in particular, circle and its inscribed quadrilateral. This is a topic that was covered in junior middle school, therefore, many students might have forgotten this topic and might not feel confident to solve it. Item 24 is related to inequality, which is normally regarded as a difficult topic for high school students (Hill, 2002). Some students did not choose any of the three questions to answer, which is not surprising in mathematics test (Jiang, Hwang, & Cai, 2002).

Rasch Analysis
The Rasch model was applied to address the first research question. Results of the principal component analysis of the residuals indicated that the unidimensionality assumption was tenable. The Rasch model assumes unidimensionality, so unidimensionality was examined with Mean Square (MNSQ) item fit statistics. The MNSQ fit values between 0.6 and 1.4 were considered reasonable (Bond & Fox, 2002). The Rasch dimension explained 57% of the variance in the data.

VALIDITY OF NHEEE 11 The largest secondary dimension accounted for only 3.7% of the unexplained variance, with an eigenvalue of 1.7. The fit statistics for all items were within acceptable limits: The infit MNSQ ranged from 0.88 to 1.02; the outfit MNSQ ranged from 0.84 to 1.75.

CONCLUSION AND DISCUSSION

This study was conducted to examine the reliability and validity evidences of the scores derived from the 2002 NHEEE mathematics test. Results of the principal component analysis of the residuals and item fit indices indicated that the unidimensionality assumption was tenable. Predictive aspects of validity were found for the 2002 NHEEE mathematics test, which suggests that these test scores are trustworthy even though the students took this test in a simulated context. However, item-level data for the 2002 NHEEE mathematics test scores were not available. Future studies should try to use the actual performance data from the NHEEE test scores rather than data collected from a simulated situation.

Analyses of the rating scale structure indicated that the original five-category rating structure did not function properly and that the three-category structure was found to be preferable. This finding potentially indicates test construction problems. As seen from the item difficulty measures, some items are relatively easy for the participants, but other items are very difficult, which means that either all students got them right or all students got them wrong. In teaching mathematics, teachers usually set up some sub-questions to help students find their solutions to difficult items, and students can get partial credits for each sub-question. Though more difficult items may increase the stress of test-takers, the use of sub-questions may alleviate the pressure and help them solve the difficult problems. This result has implications for test developers of high-stakes tests in all countries.
The hierarchy of item difficulties appeared to be consistent with the theoretical expectations, supporting evidence for construct validity for the NHEEE mathematics test. Findings indicate that the two items about set, numbers and operations are easy. There are two items about data analysis and probabilities. One is easy and the other is at the intermediate level. For the items about algebra, geometry and measurement, their difficulty levels varied from easy to the most difficult. The topic of set, numbers and operations is no longer important in high school mathematics curriculum, but the topic of data analysis and probabilities is (Chinese Ministry of Education, 2003). More items at the intermediate level about data analysis and probabilities could be included in the test. This result calls for the close match between curriculums and testing so that what is tested reflects what is taught in the classroom.

References


