Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 2 | Issue 1 Article 3

2021

Application Innovation of Educational Measurement Theory, Method, and Technology in China’s New College Entrance Examination Reform

Zhengyan Liang
South China Normal University

Minqiang Zhang
South China Normal University

Feifei Huang
South China Normal University

Derong Kang
South China Normal University

Lingling Xu
South China Normal University

Recommended Citation

Liang, Zhengyan; Zhang, Minqiang; Huang, Feifei; Kang, Derong; and Xu, Lingling (2021) "Application Innovation of Educational Measurement Theory, Method, and Technology in China’s New College Entrance Examination Reform," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*: Vol. 2 : Iss. 1 , Article 3.
Available at: https://www.ce-jeme.org/journal/vol2/iss1/3

Application Innovation of Educational Measurement Theory, Method, and Technology in China's New College Entrance Examination Reform

Zhengyan Liang [a], Minqiang Zhang [a], Feifei Huang [a], Derong Kang [a], and Lingling Xu [a]

[a] South China Normal University

Abstract

China's new college entrance examination (the *new gaokao*) reform provides an opportunity for researchers and practitioners of educational measurement to directly participate in the reform. Therefore, conducting in-depth research on the characteristics of the *new gaokao* and the issues it faces, and finding corresponding solutions theoretically, methodologically, and technically will not only help to deploy the education examination reform smoothly, but also expand and enrich the research and application of educational measurement. This article provides discussions and suggestions on some issues related to the *new gaokao* reform, including the stability issue of the examination brought by the scoring methods and subject selection, the equating issue of test scores due to biannual tests or cross-year comparisons, and the issue of giving feedback to basic education based on the analysis of the *gaokao* data.

Keywords

China's new college entrance examination reform; independent subject selection; higher-order DINA model; linear logistic test model

Introduction

China's college entrance examination, also known as *gaokao*, has a long history. Especially in the past 40 years, it has played an important role in selecting talents for colleges and universities, encouraging young people to learn, and training professional talents for the country. However, due to the highly competitive nature of gaokao and its "baton" role, schools and local education management departments regard promotion rate as the only evaluation criterion, which has had a huge impact on basic education and teaching. The society's criticism of gaokao has never stopped, and there are loud calls for its reform. In 2014, the State Council of the People's Republic of China issued the *Implementation Opinions of the State Council on Deepening the Reform of the Examination Enrollment System* (*Implementation Opinions* for short), which opened the prelude to a new round of gaokao reform. In this article, we called this the *new gaokao reform*.

New gaokao reform has brought many challenges to the traditional test development and evaluation. First, in the context of the reform, the academic achievement test is divided into the qualification test and selective test. From the perspective of educational measurement, the former is a criterion-referenced test, and the latter has both the nature of criterion- and norm-referenced tests. For the selective test, the raw score is converted into grade score (e.g., there are five grade, the top 15% is converted to A, 35% becomes B, 35% as C, 13% as D, and the last 2% is going to be E) and then counted into the total score of gaokao, which is used as the basis for admission. In test development, it is necessary to consider the coexisting properties of criterion- and norm-referenced tests in the selection of test content, item design, and the evaluation of items and tests. This requires innovations in the analysis method and the theoretical basis of examination and evaluation.

Second, the scheme of *biannual tests* (for English) puts forward new requirements for testing data analysis and score reporting. The same group of examinees take the examination twice a year, which requires the equating of scores of these two examinations. However, the anchor-item/person design and other commonly used equating methods become invalid because of the high-stakes nature of gaokao. This requires the exploration of new equating methods.

Third, the reform requires a comprehensive quality evaluation of students as the reference basis for college

CONTACT: Minqiang Zhang. 2640726401@qq.com. South China Normal University. Guangzhou, China, 510631.
Translator: Peida Zhan. pdzhan@gmail.com. Department of Psychology, College of Teacher Education, Zhejiang Normal University.

admission. However, it still lacks effective means for guaranteeing the scientific nature of the comprehensive evaluation. Therefore, it is hard to ensure the credibility of the evaluation results and the "hard link" with college admission.

Aiming at the main issues faced by new gaokao, this article provides suggestions and methods from three aspects: innovating item evaluation mode, constructing test banks and item banks, and strengthening data feedback.

1 Background and Characteristics of Traditional Gaokao

Traditional gaokao adopts separate testing modes for liberal arts and sciences. Liberal arts students are tested for Chinese, Mathematics (liberal arts), English, and comprehensive liberal arts (namely, History, Geography, and Politics). Sciences students are tested for Chinese, Mathematics (sciences), English, and comprehensive sciences (namely, Physics, Chemistry, and Biology). There is a clear boundary between liberal arts and sciences. Students need to choose their own testing mode, liberal arts or sciences, in senior high school. Then, according to students' choices, the school divides students into different classes (liberal arts classes or sciences classes) and arrange corresponding teachings. However, this kind of premature division has been criticized nationwide, as it has made liberal arts students lack the knowledge of natural sciences and sciences students lack the knowledge of humanities and social sciences.

There is a large number of gaokao examinees in China, with about 10 million examinees taking the unified examination at the same time every year. Meanwhile, gaokao is also one of the most authoritative examinations. Its scores are recognized by many countries and serve as one of the qualifications for college admissions. Therefore, it takes a lot of manpower and material resources to keep the balance between the tests and the examinees with different abilities in different regions.

In each year, the National Education Examinations Authority needs to prepare multiple sets of test forms (including test forms I, II, and III). These sets of test forms basically have the same structure, but with different difficulty gradients. Each province selects one of them according to its own characteristics.

In addition, gaokao has the dual characteristics of the provincial enrollment plan and the national unified examination. The number of examinees in different provinces varies greatly. The province with the largest number of examinees has as many as one million, while the province with the fewest examinees has only 50,000. However, because the enrollment plan is formulated on a provincial basis, the scores required by the same university recruiting examinees from different provinces are not the same.

Traditional gaokao has its advantages and disadvantages. The separate testing modes for liberal arts and sciences make the subjects fixed for examinees who are in the same mode. In addition, the college admissions adopt separate admissions of liberal arts and sciences. Thus, examinees are ranked, compared, and admitted only in their own test mode, which is relatively fair. However, due to the high-stakes nature of gaokao, schools often adopt test-oriented education to pursue a high promotion rate, resulting in the phenomenon of "no testing no teaching, and no testing no learning." This leads to students having not only incomplete knowledge structure, but also certain knowledge and ability deficiencies in professional development and interdisciplinary learning and application after entering universities. As a result, Chinese college students are often criticized for their lack of knowledge structure, innovation, development momentum, and so on, which undoubtedly causes serious obstacles to the all-round development of students.

2 Characteristics of New Gaokao

Implementation Opinions stated:

> The Examinee's total score is composed of the scores of Chinese, Mathematics, and English in the unified gaokao and the scores of three subjects in the selective examination. Specifically, keeping the Chinese, Mathematics, and English in the unified gaokao unchanged, keeping the full scores of them unchanged, and no longer separating liberal arts and sciences. Additionally, providing two test opportunities for English each year. According to the requirements of the colleges and examinees' own specialties and hobbies, the self-selected subjects to be included in the total score can be chosen from Politics, History, Geography, Physics, Chemistry, Biology, and so on.

The new gaokao reform gives examinees full freedom to choose subjects. It replaces the original testing mode of separating liberal arts and sciences with the "3+3" or

"3+1+2" mode. Specifically, Chinese, Mathematics[1], and English tests are uniformly organized by the Ministry of Education. Among the other six subjects (i.e., Physics, Chemistry, Biology, Politics, History, and Geography; some provinces also add Information Technology), examinees can choose three of them to be tested in the selective examination organized by the provinces. Figure 1 depicts the relationship between different examinations in the context of the new gaokao reform.

According to the reform requirements, the new gaokao is divided into two parts: the provincial selective examination and the national unified gaokao. The Ministry of Education is responsible for the proposition and organization of the unified gaokao in Chinese, Mathematics, and English. The results of this part will be calculated into the total score in the form of raw score (standard score in Hainan Province).

The provincial education department is responsible for the proposition and organization of the senior high school graduation test (i.e., qualifying test) and the selective examination for Physics, Chemistry, Biology, History, Geography, and Politics. The results of the selective examination are added to the total score by means of grade score (standard score in Hainan Province). It should be noted that some provinces use the "3+1+2" mode rather than the "3+3" mode. In those provinces, in addition to the three main subjects (Chinese, Mathematics, and English), the score of Physics or History is also counted into the total score in the form of raw score, and the other two selected subjects are counted in the form of grade score.

In order to avoid haggling over every score, among the 14 provinces or cities that have taken part in the reform, 13 of them use the grade score to report the results of the selective examination. Based on the total number of students and the academic performance of each province, the distribution of grades differs slightly.

3 Challenges of New Gaokao Reform to Educational Measurement Theory, Method, and Technology

Table 1 compares the differences between the traditional gaokao and new gaokao in terms of testing mode, testing subject, evaluation method, the right of subject selection, and the number of examinees. It is not difficult to find that the new gaokao gives examinees the right to choose the test subjects independently. For examinees, having the right to choose is good, but it can also be confusing because the choices are affected by many internal and external factors. Because Gaokao is a high-stakes test, examinees should not only consider their own learning interest and ability, but also strive to maximize the benefits from scores when choosing test subjects. Therefore, examinees will also consider the study schedule of each subject, the difficulty of each subject, the number of examinees taking each subject test, and the requirements of the universities and majors to be applied for. This ultimately results in uncertainty and complexity of the new gaokao reform, that is, the uncertainty in the number of examinees taking different subjects in each year, and the inconsistent distributions of examinees' abilities in different subjects.

This also presents some challenges to traditional educational measurement theories and methods when solving the above-mentioned issues faced by the new gaokao:

1. Under the new gaokao reform, the positioning of the selective examination has changed. According to the test development requirement of the selective examination, the test content now include two modules of the senior high school curriculum: compulsory and required-elective. When two modules are used as the proposition and testing requirements, the selective examination should be a criterion-referenced test, concerning whether each examinee has reached the level required for the compulsory and required-elective subjects of senior high school. However, because results of the selective examination in the new gaokao are required to be converted into grade scores and to be included in the total scores of gaokao, as one of the bases for college admissions, the selective examination then also has the function of a norm-referenced test. Thus, the proposition of the selective examination should consider the properties of criterion- and norm-referenced tests jointly with respect to test content, item design, and evaluation of items and tests. This requires the innovation of analysis methods and theories in test evaluation.

2. According to *Implementation Opinions*, in provinces that implement the new gaokao, students have two opportunities to take the senior high school academic proficiency test or English test. The scheme of biannual tests puts forward new requirements for the analysis of testing data. When the same group of examinees takes two or more examinations, there are equating issues in the requirements and design of the test form. However, due to the high-stakes nature of gaokao, the used items cannot appear again, or even similar items cannot be used. Since

[1] Mathematics is no longer divided into liberal arts mathematics and sciences mathematics.

Figure 1

The Relationship Between Different Tests in the Context of New Gaokao Reform

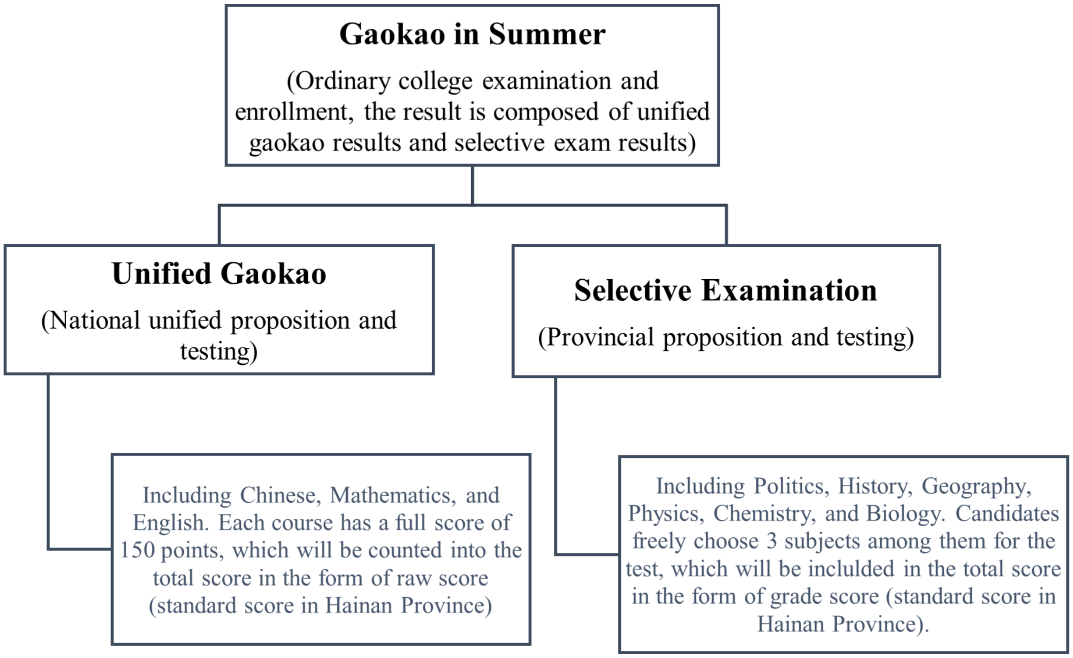

Gaokao in Summer
(Ordinary college examination and enrollment, the result is composed of unified gaokao results and selective exam results)
Unified Gaokao
(National unified proposition and testing)
Selective Examination
(Provincial proposition and testing)
Including Chinese, Mathematics, and English. Each course has a full score of 150 points, which will be counted into the total score in the form of raw score (standard score in Hainan Province)
Including Politics, History, Geography, Physics, Chemistry, and Biology. Candidates freely choose 3 subjects among them for the test, which will be inclulded in the total score in the form of grade score (standard score in Hainan Province).

Table 1

A Comparison Between Traditional Gaokao and New Gaokao

	Traditional Gaokao	**New Gaokao**
Testing mode	Division of liberal arts and sciences	No division of liberal arts and sciences
Testing subject	**Liberal arts**: Chinese, Liberal Arts Mathematics, English, History, Geography, Politics; **Science**: Chinese, Science Mathematics, English, Physics, Chemistry, Biology	Chinese, Mathematics, and English are organized by the unified national test, and the remaining six subjects (Physics, Chemistry, Biology, Politics, History, and Geography) are organized by the provinces. Each candidate can choose three subjects among them.
Evaluation method (total score) [a]	Raw score	Raw score is combined with grade score (former is used for Chinese, Mathematics, and English; latter is used for the other three selected subjects)
The right to choose subject	No	Yes
Number of examinees	The number of examinees for liberal arts and sciences is determined	The number of examinees for different combinations of selective subjects is indeterminate

[a] The scores of all courses and the total scores of Hainan Province are presented by standard scores.

the pretest cannot be carried out, the traditional anchor item/person design and other commonly used equating methods are invalided. It is necessary to explore new equating methods.

3. The reform requires the comprehensive quality evaluation of students as the reference basis for college admission. However, there is still a lack of corresponding means to ensure the scientific nature of such evaluation. At present, it is basically in the state of *soft link* with college admission, that is, the comprehensive quality evaluation has limited value for admission reference. Therefore, currently, we are still unable to find an appropriate evaluation mode or operational standard to solve the issue of how to take the data of non-academic factors (e.g., students' moral quality, physical and mental health, interests and hobbies) as the criteria for evaluating students and make them quantifiable and comparable, which is also an issue that needs to be solved by educational measurement methods.

4 Innovative Application of Educational Measurement in New Gaokao Reform

The new gaokao reform relates to numerous issues, due to the page limitation, this article only discusses the following three important issues.

4.1 Innovative Application of Item Evaluation

The new gaokao reform has the characteristics of both criterion- and norm-referenced tests. In such a case, which theory and method should be used to evaluate the items is an urgent problem to be solved. In the classical test theory (CTT), since the parameter calculation depends on the group of examinees (i.e., sample-dependent statistics), the reliability, validity, item difficulty, and item discrimination based on CTT will change with different groups of examinees. Therefore, the innovative application of item evaluation can be considered from the following two aspects.

4.1.1 Combination of CTT and Item Response Theory

In the CTT, since the parameter calculation is highly sample-dependent, the analysis of item parameters can be targeted to the group of examinees with different ability levels (i.e., stratified analysis). Specifically, examinees with different ability levels can be stratified according to different university admission batches (such as the top 10%, top 20%, top 50%, and other levels). Then, different item parameters of each group of examinees at different levels can be calculated.

Item response theory (IRT) resolves the sample-dependent problem in CTT. It can realize the cross-group invariance of item parameters and also define the latent ability parameter and item difficulty parameter on the same measurement scale. The relationship between the latent ability parameter and item difficulty parameter can be visually presented via the item characteristic curve.

Therefore, combining the analysis results of different groups of examinees with different ability levels in stratified analysis in CTT and the item characteristic curve in IRT may solve the issue of test suitability of examinee with different ability levels (i.e., the item is more suitable for examinees of which ability level). Moreover, items with different applicable scope constitute a test with both the characteristics of criterion- and the norm-referenced tests. For example, Table 2 and Figure 2 show the stratified analysis results of an item in CTT and the item characteristic curve of that item in IRT, respectively. For this item, according to the analysis in CTT, the top 10% examinees have the largest degree of discrimination. Meanwhile, according to the item characteristic curve in IRT, examinees with the latent ability value around 1.89 have the largest degree of discrimination. In brief, this item is more suitable for examinees with a high ability level. Further, this item can be labeled as a norm-referenced item, which is suitable for selecting examinees with a high academic level.

In addition, the stratified analysis of examinees at different levels is helpful for establishing the classified test, diversified admission, and independent admission. It is also of great value for establishing a comprehensive evaluation mechanism for talents and promoting scientific selection of talents.

4.1.2 Introducing a New Measurement Model: Higher-Order Cognitive Diagnosis Model

Large-scale and high-stake tests such as Gaokao are often held regularly and in a state of long-term operation. The operation process can be divided into multiple links that are relatively independent and interconnected at the same time. Smooth operation depends on good organization and coordination within and between links. However, since the selective examination in the new gaokao has the characteristics of both criterion- and the norm-referenced testing, there is still a lack of effective theories and technical to connect different links. In practice, each link is easily disconnected, or even in a completely independent state, which adds to the difficulty of test operation. By combining

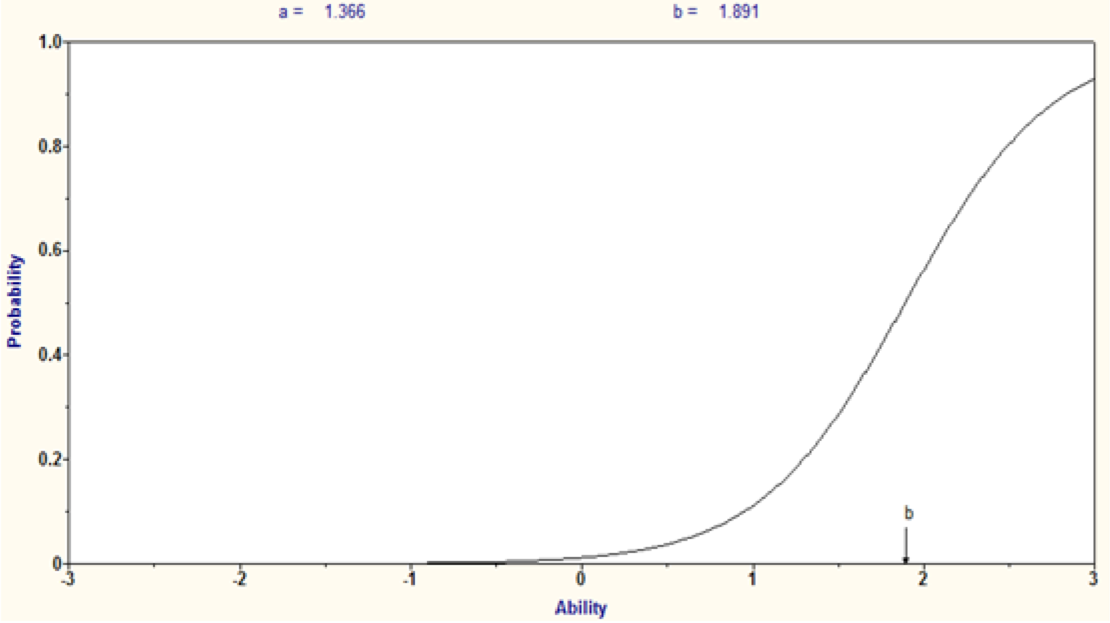
Table 2

Test Parameters of Examinee Groups with Different Ability Levels in CTT

Parameter	All	Top 2%	Top 6%	Top 10%	Top 50%	Last 40%
Difficulty	0.11	0.76	0.58	0.47	0.18	0.00
Discrimination	0.21	0.39	0.44	0.51	0.33	0.00

Figure 2

Item Characteristic Curve in IRT



IRT and cognitive diagnosis theory (CDT) together, we can develop appropriate technical methods to more effectively connect different links and facilitate test operation.

The higher-order cognitive diagnosis model (HO-CDM) to some extent integrates IRT and CDT and can provide both criterion- and norm-referenced information. Its appropriate application can not only contribute to the test design, but also facilitate the operation and evaluation of the test. The higher-order latent structure component can provide norm-referenced information. Theoretically, this component can accommodate multidimensionality and different structures; in practice, unidimensionality and common loading structure are often used. The measurement model component can provide criterion-referenced information. This component can be various reduced CDMs (e.g., DINA, DINO, LLM) or saturated CDMs (e.g., G-DINA and LCDM). In this article, the HO-DINA model (De La Torre & Douglas, 2004) is taken as an example to illustrate to modeling logic.

The HO-DINA model combines the DINA model at the lower order with the higher-order latent trait model. As a result, it can describe both the general abilities of the examinees and their mastery status of latent attributes (e.g., mastery or nonmastery). The HO-DINA model can be expressed as:

$$P(\boldsymbol{\alpha}|\theta) = \prod_{k=1}^{K} P(\alpha_k|\theta) \tag{1}$$

$$P(\alpha_k|\theta) = \frac{\exp(\lambda_{0k} + \lambda_k\theta)}{1+\exp(\lambda_{0k} + \lambda_k\theta)} \tag{2}$$

$$P(\boldsymbol{\alpha}) = \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_k \tag{3}$$

where λ_{0k} is the intercept of attribute k, and λ_k is the loading of attribute k on ability θ; $\delta_{j12\ldots K_j^*}$ are item parameters similar to the plain DINA model. The HO-DINA model can provide test information at both the higher and lower order, helping test developers, instructors, and educational researchers to better understand examinees' knowledge profiles.

4.2 Construction of Test Bank and Item Bank

To solve the issues caused by biannual tests, and to ensure the quality and stability of provincial test development, one might consider establishing a test bank and item bank.

4.2.1 Test Bank and Item Bank

Suppose an item bank is composed of all the high-quality items, and the test is composed of all the excellent items. However, the final test still may not be a good test. This is mainly because the cultural, economic, and educational levels of cities and regions in each province in China are very different. Besides, the item difficulty and item discrimination in CTT is sample-dependent. Therefore, the construction of the test bank and item bank is more in line with the actual demand of each province. At the same time, when there are enough tests to form a test bank, we can also extract some suitable items corresponding to the level of local examinees to form a special item bank.

The test bank and item bank is a system platform that takes the former as the foreground and the latter as the background. Such a system has both subject versatility and personalized test form. According to certain educational measurement theory, tests are used as the units to make propositions. The developed tests can be divided into distinct items for grouping, reorganization, fine-tuning, so as to realize the intercommunication, interconnection, and interchange function between the test bank and the item bank. A good test must match the ability levels of examinees, which needs to be pretested and comprehensively analyzed in order to receive the corresponding evaluation. According to propositional principles and standards, the test bank may contain multiple *parallel test* forms. Some items can be randomly selected from each parallel test form and combined into a pretest. In such a case, the pretest can be used as an anchor among different parallel tests, and then the parameters of each parallel test form could be obtained. Additionally, as the design of the test is compiled according to the principle of parallel test forms, the parameters in the pretest can also be transferred or corresponded to other untested tests, so as to achieve the effect of matching the tests with the ability levels of the examinees. This work needs the cooperation and participation of subject experts, educational measurement experts, and information technology experts to complete and achieve the desired results.

4.2.2 Equating Issue in Biannual Tests

The high-stakes nature of gaokao entails that the items cannot be and are impossible to be reused, so the methods of anchor item/person in equating cannot be implemented in gaokao. Therefore, combined with the actual situation of the new gaokao reform, new concepts, theories, and methods are needed to solve the equating issue in biannual tests. Previous studies have proposed new methods or ideas to replace the traditional anchor item/person methods (Mislevy, Sheehan, & Wingersky, 1993; Wright & Dorans, 1992; De Boeck & Wilson, 2004; Wei, Liu, & Dorans, 2013; Liao & Livingston, 2012; Wei & Morgan, 2016). Specifically, in the explanatory item response model (EIRM) proposed by De Boeck and Wilson (2004), the Linear Logistic Test Model (LLTM) is worth trying for realizing the idea of equating without any anchor item or person.

EIRM is proposed within the framework of a generalized linear mixture model. It is used to estimate latent ability parameters and item parameters in the IRT model through the attributes of the item and person (De Boeck & Wilson, 2004). In EIRM, LLTM estimates the attribute effect of the item rather than the individual item effect. In addition, it can also study the interaction between the item attributes by including additional parameters in the model. Unlike the Rasch model, LLTM predicts the item parameter by item attributes:

$$n_{pi} = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik} - \varepsilon_i \tag{4}$$

where X_{ik} is the score of item i on attribute k $(k = 1, \ldots, K)$; β_k is the regression weight of item attribute k. Therefore, compared with the Rasch model, LLTM uses a linear equation to estimate the item parameter β_i:

$$\beta_i{}' = \sum_{k=0}^{K} \beta_k X_{ik} + \varepsilon_i \tag{5}$$

It is worth noting that when the predictive effect of item attributes is not good enough, β_i and $\beta_i{}'$ are not equal.

The implementation process of equating based on LLTM is similar to equating based on IRT. First, based on the common attributes of the items, the response matrix and item attribute matrix are generated according to the responses of different examinees in different test forms. Second, through LLTM, the latent ability parameters and item parameters are simultaneously estimated. Finally, the equating of test score can be achieved by linking the item parameters. Note that when using concurrent calibration on the response data obtained using different test formats, the parameter estimation and scaling process are performed simultaneously. Taking the dichotomous two-parameter logistic model as an example, the relationship between the latent ability parameter and item parameters can be expressed as:

$$P(\theta) = \frac{\exp^{a_i(\theta_p - \sum_{k=0}^{K} \beta_k X_{ik} - \varepsilon_i)}}{1 + \exp^{a_i(\theta_p - \sum_{k=0}^{K} \beta_k X_{ik} - \varepsilon_i)}} \tag{6}$$

where θ_p is the latent ability of person p; a_i is the discrimination of item i; ε_i is the residual term, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$; other parameters have been defined above.

The introduction of the equating method based on LLTM is of great significance to the test equating of biannual tests.

4.3 Making Good Use of Testing Data Feedback in Basic Education and Teaching

The data of gaokao and senior high school academic performance examination are highly authoritative and scientific, but the data of the previous gaokao are only used as the basis for students' graduation or college admission. Also, a lot of useful information contained in gaokao data cannot be used because of confidentiality. To make good use of these data and fully utilize educational evaluation as a "baton," it is not recommended to seal up the data of gaokao for confidentiality reasons, but to establish laws and regulations for the use and disclosure of these data.

First, in terms of data analysis, a general technical principle of data analysis should be developed. At present, each province has its own test, and the submitted data are very random, which cannot be processed uniformly. As a result, the scores and parameters obtained by various calculations are not comparable, lacking scientificity and reliability. Therefore, it is necessary to have a unified requirement regarding the form of data submission and the rules of data analysis.

Second, in terms of the application of methods, more statistical analysis methods (e.g., hierarchical linear analysis, latent class analysis, latent profile analysis) should be introduced according to different evaluation purposes and requirements. The application and innovation of statistical analysis methods can carry out deeper and more detailed data mining, thereby obtaining more targeted conclusions and providing diagnostic feedback. Furthermore, it can help educational administrative departments at all levels to accurately grasp the current status of basic education and teaching, and help them to improve teaching level in a targeted way.

Third, the longitudinal data analysis method should be emphasized. Current educational evaluation is mostly based on cross-sectional data, and the analysis of cross-sectional data can provide the status of the examination in the current year immediately, which has certain reference value. Education is a continuous development process. To make a scientific and reasonable educational evaluation, longitudinal data analysis is needed. Longitudinal data analysis can be included in the new gaokao reform to establish the cross-year comparison of the abilities of examinees, draw attention to personal development, and serve as one of the references for evaluating the quality of reform.

Fourth, it is necessary to provide feedback services based on evaluation results of educational testing data for various stakeholders (e.g., students and teachers of the city, district, and county). The results of the educational evaluation should be addressed not only to educational decision-makers, but to all participants in educational tests. On the one hand, it is vital to let the superior departments have overall command of the situation and understand the status of education and teaching of the subordinate departments. On the other hand, it is necessary for subordinate departments to know the problems existing in education and teaching. Besides, the feedback of evaluation results can help teachers to know their own problems in the teaching process and help students to understand the mechanism of the learning process. However, because different stakeholders focus on different points, it is needed to provide more targeted evaluation results. Further, it is of great importance to combine educational evaluation with computer technology to realize the customization and intelligent feedback of evaluation results.

Fifth, it is necessary to conduct specialized research on the educational measurement of comprehensive quality evaluation. The evaluation model and platform should be developed to study the combination of and transformation between qualitative data and quantitative data, so as to find a fair, scientific, and recognized evaluation method and model. This is an innovative research of educational measurement theory and method, which needs investment and time. It depends on the joint participation and practices of schools and scientific research institutions.

China's new college entrance examination reform puts forward many new requirements for the research and application of educational measurement, and also presents great challenges to every participant. Therefore, it is vital for the researchers of educational measurement and examination evaluation to innovate theoretically and technologically on a deeper level in order to resolve the problems faced by China's new gaokao.

References

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer..

De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

Liao, C., & Livingston, S. (2012). *A search for alternatives to common-item equating.* Paper presented at the annual meeting of National Council on Measurement in Education, Vancourver, British Columbia, April.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55–78.

Wei, Y., Liu, J., & Dorans, N. J. (2013). *Evaluation and exploration of linking design for a language test.* Unpublished manuscript.

Wei, Y., & Morgan, R. (2016). *An evaluation of the single-group growth model as an alternative to common-item equating* (Research Rep. 16-01). Princeton, NJ: ETS.

Wright, N. K., & Dorans, N. J. (1992). *Using the selection variable for matching or equating* (Research Rep. 93-04). Princeton, NJ: ETS.