Song and He *Language Testing in Asia* (2015) 5:4
DOI 10.1186/s40468-014-0011-z

Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

The effect of a national education policy on language test performance: a fairness perspective

Xiaomei Song[1]* and Lianzhen He[2]

* Correspondence:
sxmdaphne@yahoo.com
[1]Georgia Southern University, Statesboro Georgia, USA
Full list of author information is available at the end of the article

Abstract

Project 211 is one of the most important educational policies in China, which aims at selecting a small number of "key universities" for sustainable development in the 21st century. These selected "key universities" have received substantial funding from the government so they can recruit outstanding faculty and be equipped with high quality facilities. Although this national policy has come into being for more than a decade, limited empirical studies have been conducted to examine how the policy influences students' performance as well as whether the policy is perceived to interfere with test fairness. Using differential item analysis (DIF) and content analysis, this study examined the effect of group membership (key vs. non-key universities) on one of the large-scale high-stakes language tests–the Graduate School Entrance English Examination (GSEEE). The results identified a number of DIF/DBF, all favoring test takers from key universities. A review of the flagged items by three test reviewers found a myriad of potential factors including unbalanced learning opportunities, which may contribute to performance differences between two groups. However, none of the reviewers deemed that unbalanced educational opportunities brought bias and presented a threat to fairness. Fairness in this context is conceptualized to require individuals or groups who are less well-off to sacrifice social benefits and resources to those who are better off to achieve the overarching aim of maximizing economic benefit to the society.

Keywords: Test fairness; Project 211; Differential item analysis; Content analysis; Utilitarianism

Background

Project 211 is one of the important national education policies in China. Launched in 1993 by the Ministry of Education of People's Republic of China, it aims at curbing the regional gap and improving the quality of higher education, scientific research, administration, and institutional efficiency as a basis for training higher-level professional manpower for China and strengthening those institutions and disciplines as a national priority in higher education (Liu 2009). According to this policy, there are two types of universities in the Chinese higher educational system: key and non-key universities. Colleges and universities are assessed by various criteria such as staffing, buildings, libraries, laboratories, research, funds, and prestige in specified disciplines to determine whether they are "qualified" to be included as top institutions (Huang 2008). After several rounds of evaluation, among 1700 colleges and universities across China, about 113 universities have been selected as key universities. Most key universities are located in the eastern economic-developed urban areas. The central government has provided tremendous funding with these selected key universities so they can be equipped with better facilities and be able to recruit outstanding teachers, scholars, and researchers in different areas and disciplines (Mok 2006).

Although this national policy has come into being for more than a decade, there have been limited empirical studies investigating how the policy influences student language test performance as well as whether this policy is perceived to interfere with the fairness of large-scale high-stakes language testing in China. In language testing, one aspect that has attracted much attention is opportunity to learn in the discussion of test fairness (Kunnan 2008). While some researchers and testing specialists dismiss it in discussing test fairness (Willingham and Cole 1997; Xi 2010), more and more researchers believe that a fair test should be accessible to all test takers in terms of learning opportunities since what test takers know and can do in any testing situations rely on the educational opportunities offered to test takers (Stobart 2005; Pullin and Haertel 2008). According to Gipps and Stobart (2009), issues of test fairness are bound to the joint consideration of various educational values and arguments in the content of historical perspectives. Considering the key role of high-stakes testing in the Chinese educational system (Berry 2011), whether and how university types interacted with language test performance has important implications in exploring fairness across a large country of China.

Using Differential Item Functioning (DIF) and content analysis, this study examined the effect of group membership (key or non-key universities) on one of the large-scale high-stakes tests – the Graduate School Entrance English Examination (GSEEE) of the 2009 administration. The study first examined whether the GSEEE items functioned differentially towards different university type groups. It then examined whether these flagged items, if any, might bring bias towards certain group based on content analysis of expert test reviewers. The GSEEE is a national, standardized, language test that measures test takers' knowledge of English and provides information for educational institutes to select candidates for their master's programs (Cheng 2008; He 2010). It is a norm-referenced test, which is applied to all the non-English major applicants in any areas of Humanities, Social Sciences, and Sciences. The GSEEE impacts over one million test takers who compete for a limited number of spaces in higher education. Only those students whose total scores on the GSEEE are above the admission cut-scores set by the country's Ministry of Education may have a chance to attend university graduate programs. According to the test specification for the GSEEE administered in 2009, the GSEEE examined test takers' linguistic knowledge in grammar and vocabulary, and skills in reading and writing. The total number of the GSEEE test takers in 2009 reached approximately 1.27 million and the acceptance rate was 32.86% (Ministry of Education, 2009). This study addressed the following two research questions:

1) How do the GSEEE items and bundles, if any, exhibit differential functioning toward test taker groups who are from key or non-key universities?
2) Do test reviewers perceive opportunity to learn due to the university type as one of the possible causes for the differentially functioning GSEEE items and bundles? If so, is opportunity to learn linked to the potential bias?

Opportunity to learn (OTL), educational policy, and test performance

An essential goal of high-stakes testing is to provide all test takers with equal, comparable opportunities to demonstrate what they know and can do. The literature in the West generally has a consensus about the fair treatment of all test takers on test design, development, administration, scoring, and score-based use for decision making (AERA, APA, & NCME, 2014; ILTA, 2004; McNamara and Ryan 2011). Increasingly, researchers also argue that fair testing needs to address inequitable distribution of educational resources and access to knowledge (Kunnan 2008) and opportunity to learn (OTL) falls within the scope of test fairness discussions (Pullin and Haertel 2008). This controversy makes it necessary to examine, in the Chinese context, whether differences in the educational experiences of different groups of students, as a result of educational policies, may cause differences in test performance, as well as whether such group performance differences are perceived to interfere with appropriate interpretation of test scores used for education and decision making. The Chinese language testing context presents a novel perspective as limited research has been conducted in this area.

The previous literature has indicated that educational polices impact OTL for different groups of students. For instance, advanced in 2001 in the U.S., No Child Left Behind (NCTB) Act mandates testing for all students in grades 3 through grade 8 each year and at least one during high school in the United States. This act has been spreading great controversies. Critics expressed concern that the policy perpetuated poverty and disadvantage and score differences, providing rich and poor schools with stark contrasts in learning environments and physical surroundings (Darling-Hammond 2000). Another example is Norway, in which egalitarian society with strong socio-democratic traditions, economic and cultural equality have been highly promoted in its political agenda (Carlson 2009). In 2003, the new government took initiatives to develop national testing focusing on pedagogical and public reporting functions. The introduction of national testing received a strong negative public reaction due to a fear that it was believed the national testing would contribute to increased differences between the rich and the poor (including score differences), which is against its socio-democratic traditions.

Substantial empirical studies documented the influence of OTL variables on students' test scores (Boscardin et al. 2005; Aguirre-Muñoz and Boscardin 2008). Wang (1998) examined four dimensions of OTL: content coverage, content exposure, content emphasis, and quality of instructional delivery. Results found that OTL variables were significant predictors of both written and hands-on test scores as well as variation in the effects of OTL attributable to differences in test format. Specifically, students who received longer engaged time on specific content in the classroom were likely to perform better on written tests than those who did not, and students who received better quality of instructional delivery were likely to perform better on the hands-on tests. Using interviews, Walpole et al. (2005) investigated urban African and Latino high school students' perceptions, test preparation, information sources, and strategies towards the college admission tests such as SAT and ACT. Findings showed that the African American and Latino students were generally lack of information and lack of resources to pay for tests and test preparation, as a result, they perceived that they did not have adequate and equal opportunity to demonstrate their knowledge and perform at their best.

Differences in OTL and test performance have repeatedly been reported to be associated with learner characteristics such as language background and disability. Abedi (2005) illustrated the challenges of assuring fair and valid testing for the English language learner (ELL) population under the current legislation of NCLB. Using SIBTEST, Finch, Barton, and Meyer (2009) examined whether a large-scale high-stakes test measured the construct of interest equally well for all test takers receiving accommodations and those not receiving accommodations. The study found accommodations appeared to have been detrimental in performance on certain type of test items, perhaps placing too great a cognitive burden on accommodated test takers. Kong (2009) investigated the effect of test taker geographic location on the reading comprehension section of the Test for English Majors Band 4 (TEM-4) in China. A total of eight universities located in the eastern economic-developed area and western underdeveloped area (4 each) were selected. It was unclear on what criteria those universities were chosen for the two categories. T-tests showed no significant differences in reading comprehension between test takers from developed and less developed areas. Using SIBTEST, the study found one item with C-level DIF favoring the developed area and no DBF was found. Expert review with the DIF items pointed out the potential reason for DIF existence might be related with vocabulary knowledge. The study, therefore, concluded that no bias existed with the TEM-4.

Differential item functioning

Differential item functioning (DIF) explores whether test items function differentially across different groups of test takers who are matched on ability. DIF exists when groups of test takers with equal ability have differing response probabilities of either a) successfully answering an item (i.e., in multiple choice) or b) receiving the same item score (i.e., in performance assessment) (Zumbo, 2007). The general cause of DIF is that test items measure “at least one secondary dimension in addition to the primary dimension the item is intended to measure” (Roussos and Stout, 2004, p.108). Secondary dimensions are further categorized as either auxiliary dimensions that are part of the construct intended to be measured or nuisance dimensions that are not intended to be measured. Bias, thus, might occur if the existence of DIF is due to the situation that test items measure nuisance dimensions that are not relevant to the underlying ability of interest. One of the approaches to DIF detection is the traditional, exploratory approach which is conducted in two steps: statistical identification of items that favour particular groups followed by a substantive review of potentially biased items to locate the sources of DIF (Gierl, 2005). This traditional, exploratory approach has been widely used in the previous empirical studies (Ferne and Rupp, 2007; Geranpayeh and Kunnan 2007).

To conduct the first step, several statistical procedures have been developed, including the Mantel-Haenszel method (MH), logistic regression (LR), the standardization procedure, and IRT (see a review by Clauser and Mazor, 1998). Developed by Shealy and Stout (1993), Simultaneous Item Bias test (SIBTEST) is a nonparametric procedure to estimate DIF in an item or bundle of items. Test takers are compared based on their membership in either the reference or focal group (e.g., test takers from key or non-key universities), where the suspicion is that the focal group might be disadvantaged on test items due to DIF. Items (bundles) on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest consists of those items suspected of measuring the primary and secondary dimensions; and the matching subtest contains items believed to measure only the primary dimension.

Fundamentally, SIBTEST examines the ratio of the weighted difference in proportion correct (for reference and focal group member) to its standard error. DIF occurs 1) if an item is sensitive to both the primary dimension and a secondary dimension and 2) if the reference and focal groups that have been equated on the primary dimension differ in distribution on a secondary dimension (Rousos and Stout 1996a). The SIBTEST procedure can be used to determine the extent of the DIF, and classify items as having either negligible (A-level) DIF, moderate (B-level) DIF, or large (C-level) DIF (Roussos and Stout 1996b).

SIBTEST has become one of the more popular DIF procedures. First, SIBTEST has been proven to be a powerful DIF procedure (Penfield and Lam 2000). Zheng, Gierl, and Cui (2007) investigated the consistencies and effect size of three DIF procedures: MH, SIBTEST, and LR. Results showed consistent estimates on the magnitude and direction of DIF among the three DIF procedures. Second, SIBTEST uses a regression estimate of the true score based on iterative purification instead of an observed score as the matching variable. As a result, test takers are matched on an estimated ability score rather than an observed score, which increases the accuracy of the matching variable. Third, SIBTEST can be used to explore differential functioning at the item and bundle levels. Since SIBTEST is one of a few procedures that can evaluate bundle DIF (DBF), it provides increased power through more effectively controlled Type I error. Items with small but systematic DIF may very often go statistically unnoticed, but when combined at the bundle level, DIF may be detected (Roznowski and Reith 1999; Takala and Kaftandjieva 2000). Examining DBF becomes necessary to completely understand the influence of grouping variables on test performance, especially when important, although perhaps subtle, secondary dimensions associated with different bundles have been found in tests (Douglas et al. 1996).

The substantive analysis is then conducted after the statistical DIF analysis. While DIF analyses identify differential performance across items, substantive analyses are required to determine the likely causes of the DIF and whether those causes are connected with the potential bias. The substantive analysis usually involves item reviews by subject-area experts (e.g., curriculum specialists or item writers) in an attempt to interpret the factors that may contribute to differential performance between specific groups of test takers. A DIF item is potentially biased when reviewers identify the DIF sources that are due to components irrelevant to the construct measured by the test, placing one group of test takers at a disadvantage. Exploratory DIF analyses have been widely used in previous empirical studies, despite the situation that content analysis may not always provide conclusive answers regarding DIF sources and test reviewers cannot determine decisively that the existence of DIF and DBF is due to bias (Geranpayeh and Kunnan 2007; Uiterwijk and Vallen 2005).

Methods

Subjects

The study used the GSEEE item-level data from one major university in Southern China. Using data of the 2009 administration, the applicants' background information and their GSEEE item scores were collected through one of the provincial NEEA branches. Data from a random stratified sample of 13,745 applicants (test takers) were obtained, with 57.5% of the test takers being male and 42.5%, female. Approximately 8.4% of the test takers studied in the Humanities (e.g., literature, history, and philosophy), 16.3% in the Social Sciences (e.g., economics, psychology, and management), and 75.3% in the Natural and Applied Sciences (e.g., physics, chemistry, biology, and computer sciences). Such information was similar to the demographic information of the whole school and the overall GSEEE testing population (Ministry of Education 2009).

The 2009 GSEEE included three sections (see Table 1). Section I, Cloze, was a multiple-choice (MC) test of vocabulary and grammar with 20 blanks in the text1. There were three parts in Section II, Reading comprehension (RC). Part A contained 20 MC reading comprehension items based on four reading passages on different topics; Part B was a text with five gaps where sentences were removed and test takers were required to choose the most suitable option for each gap; and Part C was a text in which five sentences were required to be translated from English into Chinese. Section III, Writing, included two parts. Part A was a practical writing task and Part B was an essay writing task.

The Section I Cloze items (1–20) and the items in Parts A and B in Section II (21–45), six texts in total, were dichotomously scored and weighted as 60 points out of a total of 100. The remaining three texts were polytomously scored. Five sentences in the text of Section II Part C translation were scored based on the overall meaning, structure, and correctness of Chinese spelling. Using the negative (error) deduction approach, each mistake based on those three criteria was penalised with 0.5 point until the total score for that sentence was 0. For marking the two writing pieces, there were six scoring criteria in the context of analytic rating rubric (Category 0–5). The “5” category, for example, is given to test takers who produce well- organized and well-developed text, address all major elements of the task, demonstrate syntactic variety and range of vocabulary, use accurate word choice and proper grammar, and display appropriate choices of forms and registry.

Analyses

Descriptive statistics were calculated to provide an overall picture of the GSEEE data set, and Cronbach’s alpha coefficients were calculated for the entire test and subtests to

Table 1 Description of the GSEEE administered in 2009

Section	Part and item	Topic	Format	Score
I Cloze	Text (Items 1–20)	Animal intelligence	MC	10
II Reading	Part A Text 1 (Items 21–25)	Habits	MC	10
	Part A Text 2 (Items 26–30)	Genetic testing	MC	10
	Part A Text 3 (Items 31–35)	Education and economic growth	MC	10
	Part A Text 4 (Items 36–40)	The history of the New World	MC	10
	Part B Text (Items 41–45)	Theories of culture	Multiple matching	10
	Part C Text (Items 46–50)	The value of education	Translation	10
III Writing	Part A	White pollution	Practical Writing	10
	Part B	Closeness and remoteness of Internet	Essay Writing	20
Total				100

provide an estimate of the internal consistency of the GSEEE. After that, the two-step exploratory approach was conducted. SIBTEST were first used to identify the presence of DIF and DBF, followed by content analysis that explored the likely causes of DIF and DBF in terms of test taker groups of university types.

SIBTEST

This current study used test takers from key universities as the focal group and from non-key universities as the reference group. SIBTEST was used for 45 dichotomously-scored items and Poly-SIBTEST was used for 3 polytomously-scored items. SIBTEST was conducted with 45 dichotomous-scored items at the both item and bundle level. A standard one-item-at-a-time DIF analysis was performed in which each item was used as a suspect item and the rest serving as the matching criterion. Items displaying DIF were then removed from the matching criterion and DIF analysis was re-conducted. After that, DBF analysis was performed. Since all dichotomously-scored items were embedded in six texts, the study examined DBF at the text level as apparently each text shared a common content theme. This bundling method is consistent with the previous literature (Douglas et al. 1996; Gierl 2005).

The test takers of the entire pool of 13,745 test takers were randomly reduced to 2000 for each group. The reference and focal group had the same number of test takers. In addition, in order to guard against unrepresentativeness within each group, we used an equal number of test takers with different characteristics of gender to facilitate comparisons. In other words, when examining effects of key universities on the GSEEE, a stratified sample of 1000 female test takers from key universities and 1000 male test takers from non-key universities were selected as the focal group; and a sample of 1000 female test takers from non-key universities and 1000 male test takers from non-key universities were selected as the reference group. This type of stratified random sampling allows us to examine group effects with test takers from a diverse spectrum of characteristics and capture the major variations between the examined groups. DIF and DBF results were validated by multiple rounds of sampling with reference and focal groups.

Content analysis

Content analysis was employed to identify the likely causes of DIF and DBF. It also examined whether the test reviewers perceived that those possible causes were linked to the potential bias toward groups of different university types. The expectation of the content analysis was that if test reviewers thought the differential functioning of those flagged items/texts was due to components irrelevant to the construct measured by the GSEEE such as opportunity to learn, then it might be possible to conclude those items may be biased. To complete the content analysis, recorded telephone interviews were conducted with three test reviewers. Three test reviewers were current university professors with extensive teaching experience in both undergraduate and graduate programs (see Table 2). The reviewers were purposely chosen based on their gender, age, and extensive knowledge of English teaching and testing. Since individual reviewers with different backgrounds could be expected to interpret and approach each DIF/DBF in different ways, this will result in a more comprehensive understanding of these flagged test items/texts.

Table 2 Background information of content reviewers

Reviewer	A	B	C
Gender	**Female**	**Female**	**Male**
Age	**46-50**	**40-45**	**51-55**
Education	**PhD**	**PhD**	**M. A.**
Professional experience	❖ 25 years of teaching experience in English ❖ Involvement in high-stakes item writing ❖ Language testing researcher	❖ More than 20 years' teaching experience in English ❖Involvement in high-stakes item writing ❖ Language testing researcher	❖ 24 years of teaching experience in English ❖ Involvement in high-stakes item writing ❖ ESL researcher

The format of the content analysis was similar to that conducted by Geranpayeh and Kunnan (2007). First of all, three participants were asked to decide whether the flagged items/texts were likely to advantage/disadvantage test takers who were from key or non-key universities. Second, they were asked to rate the suitability of the flagged items/texts based on a scale from 1 (strongly disadvantage) to 2 (slightly disadvantage) to 3 (neither advantage nor disadvantage) to 4 (slightly advantage) to 5 (strongly advantage). Third, the test reviewers were asked to explain their rating choices and make comments related to their choices. Before conducting the content analysis, the reviewers were briefed about the nature of the study, and they were given a copy of the testing paper and the items/texts needed for the content analysis.

Results and discussion

Descriptive statistics

Table 3 reports the mean scores, standard deviation, skewness, and kurtosis for each group and overall. The descriptive statistics showed that, overall speaking, test takers from key universities performed better than test takers from non-key universities. Skewness and kurtosis values ranged between +1 and −1, indicating that the distribution of the data could be considered normal. Using One-way ANOVA, it was found that there were significant differences in overall test scores between test takers from key and key universities [F (1, 13360) = 7.46, $p < .01$]. This result indicated the score differences in the GSEEE existed regarding university type groups; however, it was unclear whether the differences were caused by different proficiency ability or other reasons. Cronbach's alpha with each section and the total scores were calculated (0.53 for Section I; . .61 for Section II; . =0.65 for Section III; and . =0.71 for total). In general, these reliability estimates were not very high.

Due to the low coefficient estimates, a follow-up investigation was conducted to examine item qualities by using IRT-Bilog index. Generally speaking, the test showed a wide span of item difficulty with P-values (proportion correct) ranging from .09 to .85.

Table 3 Results of descriptive statistics

	N	Mean	SD	Kurtosis	Skewness	F	Sig.
Key university	4630	51.02	10.59	.53	-.52	372.66	*P < .01*
Non-key university	8732	47.23	10.91	.06	-.41		
Total	13362	48.55	10.96	.17	-.41		

However, item discrimination values were found generally low ranging from .02 to .35, with a large number below .20 (29 out of 45 MC items). These low values showed that the GSEEE test items did not function well to differentiate the high performers from low performers. In addition, two items– Item 12 and 43 had negative item discrimination values (−.07 and -.04 respectively). The results indicated the existence of flawed items in the 2009 GSEEE. Although the NEEA claims to have established a quality control system and conducted test evaluation research (Liu 2010), the results identified significant quality issues, which weaken the GSEEE's fairness claim.

SIBTEST

Table 4 provides an overall description of the SIBTEST results at the item and bundle (text) level. To examine whether the test quality may have had an impact on the DIF results, the DIF/DBF analysis was conducted with and without the two test items which showed negative discrimination. Results found that the quantity and size of the flagged bundles remained even after excluding these two items.

Using test takers from key universities as the focal group and non-key universities as the reference group, SIBTEST showed four flagged items/bundles, all favoring test takers from key universities at the large C level. The Beta-uni statistic was used as an effect size for gauging the magnitude of DIF. Positive Beta-Uni indicats DIF favoring the reference group while negative Bete-Uni means DIF favoring the focal group. The Cloze text (1–20) regarding animal intelligence in Section I, Text 2 (36–40) regarding the history of New World in Section II Part A, and Text (41–45) in Part B regarding theories of culture favored test takers from key universities significantly. The section of translation in Section II Part C which asks test takers to read a text about the value of education and translate some underlined sentences into Chinese also exhibited C-level DIF.

SIBTEST was used to examine whether the GSEEE items/texts functioned differentially towards test takers from different university types. SIBTEST quantified the size of DIF at the item and bundle (text) level. When discussing these SIBTEST findings, it is important to keep in mind that differences do not reflect absolute group differences but rather relative performance discrepancies regarding items/texts after the groups have been matched for overall score. As is evident from this investigation, test takers from key universities performed significantly better on certain items/texts than those from non-key universities who were matched on overall scores. Alternatively, the matched-ability groups based on overall scores had differential probabilities of success on answering the flagged items/texts, and test takers from key universities persistently

Table 4 Results of the SIBTEST analysis

Section	Item/Bundle	Beta-Uni with/without	Favouring
I Cloze	Item (1–20)	-.238/-.238	Key University
II Reading	Part A Text 4 (36–40)	-.099/-.099	Key University
	Part B Text (41–45)	-.090/-.090	Key University
	Translation	-.115/-.115	Key University

Note. p < .05. For each item/bundle, the matching subtest consisted of the remaining items/bundles with the exception of items/bundles displaying B- and C-level DIF/DBF.

outperformed those from non-key universities on the flagged items/texts. Results identified four DIF/DBF, and all of them favored test takers from key universities, indicating the effectiveness of *Project 211* in enhancing learning outcomes of students who come from key universities. The results are consistent with the literature, which highlights the disparities (e.g., geographic and urban) in learning outcomes and accessing educational resources resulting from *Project 211* (Huang 2008; Jiang and Li 2008). Since limited resources of Chinese higher education are allocated by large quantities in the key universities in the Easters coastal areas, test takers from key/non-key universities may obtain distinctive learning opportunities, in terms of both quality and quantity, which impact their test performance.

Content analysis

Three test reviewers examined whether the educational policy *Project 211* and OTL contributed to differential functioning of the items/texts identified by SIBTEST. They explored a variety of potential reasons for differential functioning, considering both test and learner characteristics. In terms of Cloze (Item 1–20), three test reviewers concluded that students from key universities performed better than those from non-key universities in that students from key universities generally had better mastery of grammatical knowledge, vocabulary knowledge, and also knowledge of language as a whole. In addition, since test takers from key universities had received high teaching qualities, learning facilities, and learning environments, they were supposed and expected to outperform those from non-key universities.

Section II Part A Text 4 (36–40) illustrates the history of new Englanders. None of the three test reviewers rated this item as unfair towards test takers from key/non-key universities. The reasons for differential functioning of this text included differences in IQ, education background in primary and undergraduate education as well as English language competency.

Section II Part B (41–45) discussed the biological evolution and analyzed which sentence fitted in the text. The primary focus of this text was to analyze and understand logical relationship within the text and paragraphs. The test reviewers felt that the major reason for group performance differences was their discrepancies in knowledge and skills. Students from key universities generally read more globally and they were better in their reasoning or logical thinking. These students performed better also because of their learning opportunities. Test takers from key universities were provided with qualified teachers and learning resources to understand the logical relationships between paragraphs in the text, especially structure.

Section II Part C (46–50) asked test takers to translate the underlined sentences into Chinese with the topic of the value of education. The three reviewers pointed out that translation was usually one of the most difficult texts in the GSEEE. Compared with test takers from non-key universities, these who studied in key universities were generally better in cognitive and logical thinking, and they received more advantaged teaching and learning resources. The reviewers claimed that, as a result, test takers from key universities were in a much better advantaged position, which was what *Project 911* aimed to.

Overall, a qualitative content review of the flagged items/texts by three reviewers did not find any evidence that these flagged items/texts exhibited potential bias towards the test taker groups. Based on the results of content analysis conducted with the reviewers, a multitude of factors have been identified that make some items/texts easier or harder for groups from key or non-key university background, including item difficulty level, general IQ, English language competency, and primary education background as well as OTL in undergraduate education. Although OTL due to the university type was one of the possible causes for the differentially functioning GSEEE items and bundles, OTL was not linked to the potential bias. None of the reviewers believed that unequal OTL led to test bias. The three reviewers concluded the existence of DIF as item impact and no test bias existed. The fundamental reason for performance differences is that test takers from key and non-key universities actually belong to different language proficiency groups. DIF may be attributed to item impact, which reflects actual knowledge differences on the construct of interest (Clauser and Mazor, 1998). Moreover, the reviewers stated that, compared with those from non-key universities, test takers from key universities were supposed and expected to have greater advantages in teaching and learning and perform better in language testing, which were what *Project 211* aimed to. This view appears to be different from the standpoint among educators and researchers in the west—OTL is a fairness issue and may threaten test fairness, especially when an authority provides differential learning opportunities among students in classroom learning (AERA, APA, & NCME, 1999; Kunnan, 2008). There is a consensus among the three Chinese reviewers that test takers from key universities are supposed to be treated differently and receive quality resources and affluent learning opportunities based on their intelligence, ability, and skills.

The results from this study confirm the culturally-embedded conceptualization of test fairness. As pointed out by Gipps and Stobart (2009; Stobart, 2005), what constitutes test fairness is situated in broad social-economic and historical contexts and is mediated by local socio-cultural perspectives and constraints. In essence, the results of this study show the reviewers' acceptance of utilitarianism, meaning that the principle of fairness in distributing educational resources is to achieve the overarching aim of maximizing economic benefit to the society (Howe, 1994). Such philosophical and epistemological principle encourages educational stratification and elite school systems, which have long been practiced to stratify those who are able to produce the most from those who are not. Actually, the "key school" scheme was introduced to secondary education in 1953, higher education in 1954, and primary education in 1962 (Gang 1996). Despite discontinuity during the Cultural Revolution (1966–1976) when most of school activities were cancelled, the "key school" system continues to remain in secondary and higher education. The main purpose of the "key school" scheme (including *Project 211*) is to give a small number of schools, colleges, and universities priority in allocating limited human and material resources, so that the training of the needed top-level manpower for China's development could be carried out more efficiently (You 2007; Yuan 1999). To determine who can enter these key schools and universities, large-scale high-stakes testing plays a key, predominant role to demonstrate student ability, classify performance groups, measure potential, and evaluate effectiveness, in order to make the most of limited educational resources. The main function of testing is selection to maximize the level of talent as an outcome. Within this context, test fairness, and broadly speaking social fairness, entail that certain individuals or groups who are less well off sacrifice social benefits and resources to those who are better off.

This philosophical stance is distinctively different from the view of egalitarianism as hold by educators and scholars in the West, which emphasizes equality in political, economics, educational, or social life (Condron 2011; Gordon 1999; Holtug and Lippert-Rasmussen 2007). As discussed earlier, in the United States, No Child Left Behind (NCTB) which mandates testing for all students in grades 3 through grade 8 has been spreading controversies. Educators in the United States express their concerns that the implementation of NCTB would reinforce the wide inequalities in income among families, with the most resources being spent on children from the wealthiest communities and the fewest on the children of the poor, especially in high-minority communities (Darling-Hammond 2000). Carlson (2009) described that assessment in its educational system in Sweden has a role to play in achieving a fair distribution of privileges, requiring where the extra resources are given to those who are in need. Therefore, testing and assessment is constructed to discriminate between the weak students and the others, but not between the average and the clever ones. However, the new policy and national testing started in 2003 differentiated the average and the clever, and eroded the egalitarianism tradition.

Recently, more and more Chinese researchers and educators raise concerns about unequal OTL and unbalanced development in teacher and staff development, teaching resources, learning materials, and school activities and involvement for disadvantaged groups and individuals (Hong 2004; Wang 2011). Physically disabled students, low socio-economic urban groups, and marginalized groups (migrant workers moving from less developed areas to developed areas), as pointed out in the literature (Jacob 2006), are largely neglected in the discussion of fairness. This unbalanced situation is highlighted in primary schools located in rural areas of China, where exits many run-down schools, inadequately prepared teachers, unattractive teaching materials, inefficient school management, and high dropout and repetition rates (Postiglione, 2006). As there exist a wide variation in a country as large and populated as China, how to balance concerns for group performance parity as well as institutional and societal benefits presents a challenge to the whole society. This complexity stems from the complexity of social values that creates tensions and the need for tradeoffs in pursuing educational goals in establishing the fairness of assessments that support those pursuits.

Conclusions

Given the significant role high-stakes testing plays, it is important to examine how tests function and what they really measures. Results of this study identified four items/texts functioned differentially at C level all advantaging test takers from key universities. The three reviewers concluded the existence of DIF as item impact and no test bias existed. Unequal opportunity to learn was not perceived to be linked to test bias. None of the test reviewers deemed that unbalanced distribution in educational resources threatens test fairness. In essence, these results reflect the philosophical view of utilitarianism and highlight the principle of fairness in distributing educational resources is to maximize economic benefit and productivity.

The study had important theory and practical implications. First, the study indicated that differential item functioning and bias/unfairness are different concepts. While DIF is a statistical procedure to examine whether test items function differentially toward different groups of test takers who are matched on ability, fairness has a strong, subjective nature and is shaped by the particular social, economic, and political context where the test operates. As Kunnan (2008) argues, fairness investigations endorse the collection of academic and professional practices, and social and political considerations of a community depending on the particular local testing situations. Second, this study also has important practical implications for the GSEEE testing practices. The study shows the urgency to improve the item quality of the GSEEE. Findings in terms of reliability and discrimination indicate limitations in the GSEEE item design and the quality control problems. Given the low reliability and discrimination values, it is of paramount significance to ensure test quality so that test takers are provided with fair, adequate opportunities to perform. It is unclear how the test items with poor quality were addressed in the score report. As large-scale high-stakes language tests in China including the GSEEE have rarely been screened for item bias (Fan and Jin 2012), the paper calls for moderation panels to conduct on-going technical examinations and review draft test materials in a systematic manner.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors read and approved the final manuscript.

Author details

[1]Georgia Southern University, Statesboro Georgia, USA. [2]Zhe Jiang University, Hangzhou, Peoples Republic of China.

Received: 13 October 2014 Accepted: 15 December 2014
Published online: 15 February 2015

References

Abedi, J. (2005). Issues and consequences for English language learners. In JL Herman & EH Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 175–198). Malden, MA: Blackwell.

Aguirre-Munoz, Z, & Boscardin, CK. (2008). Opportunity to learn and English learner achievement: is increased content exposure beneficial? *Journal of Latinos and Education, 7*, 186–205.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Berry, R. (2011). Educational Assessment in Mainland China, Hong Kong and Taiwan. In R Berry & B Adamson (Eds.), *Assessment reform in education: Policy and practice* (pp. 49–61). Dordrecht, Netherlands: Springer.

Boscardin, CK, Aguirre-Munoz, Z, Stoker, G, Kim, J, Kim, M, & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and Algebra assessments. *Educational Assessment, 10*, 307–332.

Carlson, C. (2009). Crossing the bridge from the other side: the impact of society on testing. In L Taylor & C Wei (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 344–57). Cambridge, U.K: Cambridge University Press.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*, 15–38.

Clauser, BE, & Mazor, KM. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues & Practice, 17*, 31–44.

Condron, D. (2011). Egalitarianism and Educational Excellence: Compatible Goals for Affluent Societies? *Educational Researcher, 40*, 47–55.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives, 8*(1), 1–12.

Douglas, J, Roussos, L, & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465–484.

Fan, J, & Jin, Y. (2012). *Developing a code of practice for EFL testing in China: A data-based approach*. Princeton, NJ: Paper presented in LTRC.

Ferne, T, & Rupp, A. (2007). A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. *Language Assessment Quarterly: An International Journal, 4*, 113–148.

Finch, H, Barton, K, & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment, 14*, 38–56.

Gang, W. (1996). Profiles of educational assessment systems world-wide: Educational Assessment in China. *Assessment in Education, 3*, 75–88.

Geranpayeh, A, & Kunnan, AJ. (2007). Differential item functioning in terms of age in the Certificate in Advanced English Examination. *Language Assessment Quarterly, 4*, 190–222.

Gierl, MJ. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*, 3–14.

Gipps, C, & Stobart, G. (2009). Fairness in assessment. In C Wyatt-Smith & J Cumming (Eds.), *Educational assessment in 21st century: Connecting theory and practice* (pp. 105–118). Netherlands: Springer Science + Business Media.

Gordon, EW. (1999). *Education and justice: A view from the back of the bus*. New York: Teachers College Press.

He, L. (2010). The Graduate School English Entrance Examination. In L Cheng & A Curtis (Eds.), *English Language Assessment and the Chinese Learner* (pp. 145–157). New York: Routledge.

Holtug, N, & Lippert-Rasmussen, K. (2007). *Egalitarianism: New Essays on the Nature and Value of Equality*. Oxford, UK: Oxford University Press.

Hong, S. (2004). Access to higher education for disadvantaged groups in China. *Chinese Education and Society, 37*, 54–71.

Howe, K. (1994). Standards, assessment, and equality of educational opportunity. *Educational Researcher, 23*, 27–32.

Huang, Y. (2008). On the optimization of the structure of management organization for the establishment of the key subject of the "211 project" in higher education institutions. *University Educational Science Research, 110*, 44–48.

Jacob, WJ. (2006). Social justice in Chinese higher education: Regional issues of equity and access. *Review of Education, 52*, 149–169.

Jiang, C, & Li, S. (2008). An empirical study of the distributional changes in higher education among east, middle and west China. *Frontiers of Education in China, 3*, 192–224.

Kong, W. (2009). TEM-4 yuedu ceshi de DIF yanjou [DIF Study of Reading Module in TEM-4]. *Foreign Language in China, 6*, 15–18.

Kunnan, AJ. (2008). Towards a model of test evaluation: Using the test fairness and wider context frameworks. In L Taylor & C Weir (Eds.), *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229–251). Cambridge, U.K.: Cambridge University Press.

Liu, B. (2009). The review and prospect of key university construction since the foundation of People's Republic of China. *Hebei Academic Journal, 29*(4), 1–6.

Liu, Q. (2010). The National Educational Examinations Authority and its English Language Tests. In L Cheng & A Curtis (Eds.), *English Language Assessment and the Chinese Learner* (pp. 29–43). New York: Routledge.

McNamara, T, & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*, 161–78.

Ministry of Education (2009). 2009年国家研究生招生计划. [Enrollment Planning Objectives of 2009]. Retrieved on April 20, 2011 from http://kaoyan.eol.cn/html/ky/2009kz/.

Mok, KH. (2006). *Education reform and education policy in East Asia*. New York: Routledge.

Penfield, RD, & Lam, TCM. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and practices, 19*(3), 5–15.

Postiglione, GA. (2006). School and inequality in China. In GA Postiglione (Ed.), *Educational and social change in China: Inequality in a market economy* (pp. 3–24). NY: M.E. Sharpe, Inc.

Pullin, DC, & Haertel, EH. (2008). Assessment through the lens of "Opportunity to Learn". In PA Moss, DC Pullin, JP Gee, EH Haertel, & LJ Young (Eds.), *Assessment, Equity, and Opportunity to Learn* (pp. 17–41). Cambridge, UK: Cambridge University Press.

Roussos, LA, & Stout, WF. (1996a). A multi-dimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355–371.

Roussos, LA, & Stout, WF. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Roussos, LA, & Stout, WF. (2004). Differential item functioning analysis. In D Kaplan (Ed.), *The Sage handbook for social sciences* (pp. 107–115). Newbury Park, CA: Sage.

Roznowski, M, & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*, 248–269.

Shealy, R, & Stout, WF. (1993). A model-based standaradiation approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Stobart, G. (2005). Fairness in multicultural assessment. *Assessment in Education: Principle, Policies, and Practices, 12*, 275–87.

Takala, S, & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing, 17*, 323–340.

Uiterwijk, H, & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 22*, 211–34.

Walpole, M, McDonough, PM, Bauer, CJ, Gibson, S, Kanyi, K, & Toliver, R. (2005). This test is unfair: Urban African American and Latino high school students' perceptions of standardized college admission tests. *Urban Education, 40*, 321–49.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*, 137–156.

Wang, H. (2011). Research on the influence of College Entrance Examination policies on the fairness of higher education. *Chinese Education and Society, 43*, 15–35.

Willingham, W, & Cole, NS. (1997). *Gender and fair assessment*. Mahwah, New Jersey: Lawrence Erlbaum.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*, 147–170.

You, Y. (2007). A deep reflection on the "key school system" in basic education in China. *Frontiers of Education in China, 2*(2), 229–239.

Yuan, Z. (1999). *On Chinese educational policy transformation: Case studies on equality and efficiency of key-point middle schools in China*. Guangzhou, China: Guangdong Educational Press.

Zheng, Y, Gierl, MJ, & Cui, Y. (2007). *Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and Logistic Regression procedures*. Chicago, IL: Paper presented at the annual meeting of the NCME.

Zumbo, BD. (2007). Three generation of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly: An International Journal., 4*, 223–233.