TEST FAIRNESS IN A LARGE-SCALE HIGH-STAKES LANGUAGE TEST

by

XIAOMEI SONG

A thesis submitted to the Faculty of Education

In conformity with the requirements for

the degree of Doctor of Philosophy

Queen’s University

Kingston, Ontario, Canada

June, 2014

Copyright ©Xiaomei Song, 2014
Abstract

Inquiry into fairness of a test has been recognized as an important research activity to direct efforts to reduce bias and discrimination against certain groups of test takers, create equal opportunities for test takers to demonstrate their knowledge and skills, and promote social justice. Given the importance of fairness in testing, my research examined the fairness of one large-scale high-stakes test in China—the Graduate School Entrance English Exam (GSEEE). To guide this doctoral research, I first drew on four sources of conceptual and empirical work to identify key issues encompassing test fairness. Informed by Willingham’s conceptual framework (Willingham & Cole, 1997; Willingham, 1999), I investigated the fairness of the GSEEE in two studies. In Study 1, I examined whether the GSEEE test items functioned differentially and brought potential bias towards test taker groups based on gender and academic background. In addition to the results of low reliability and discrimination values, Study 1 found a number of items/texts functioned differentially towards certain test taker groups. However, a qualitative content review of the flagged items/texts by three reviewers did not find consistent results regarding whether these flagged items/texts exhibited potential bias towards the test takers groups.

In Study 2, I investigated perceptions of the fairness of the GSEEE as expressed by program administrators, teachers, and test takers. While some practices were perceived as being fair, some were considered to undermine fairness. Additionally, divergences existed regarding the perceptions of the fairness of some practices. Despite mixed perceptions, the three groups of the GSEEE stakeholders perceived that, overall, the GSEEE was a fair test within the Chinese testing context. Such overall perception
was linked to reasons in two areas: the fair testing process prioritizing equality as well as the shared beliefs and priorities that highlighted equality among the participants in the Chinese testing context.

In conclusion, this research offers empirical information with regard to the fairness of the GSEEE from psychometric and stakeholder perspectives. The research also provides evidence that the conceptualization of test fairness is mediated by contextualized beliefs and traditions. Whether a test is perceived as fair or not is derived from considerations in both the testing process and the broad socio-cultural context.
Co-Authorship

Xiaomei Song is the PhD candidate. This manuscript-method dissertation was developed with guidance from her academic supervisor Dr. Liying Cheng and supervisory committee Drs. Don Klinger and Benjamin Kutsyuruba. Xiaomei Song conducted all of the literature reviews, research design, data collection, data analyses, and writing for this dissertation. In conducting Study 1, Dr. Cheng assisted with the access to the dataset and Dr. Klinger offered guidance in the statistical analysis. Drs. Cheng and Klinger also contributed to Study 1 by editing of the manuscript in preparation for journal publication. They are co-authors on this manuscript whereas Xiaomei Song is the main author. Overall, Xiaomei Song took full responsibility for the materials in this dissertation.
Acknowledgements

The dissertation provokes my thoughts not only on test fairness but also on
fairness in general that seems to be an omnipresent topic in our life. During the pursuit of
this research topic, I would like to express my great gratitude to my Ph.D. supervisor, Dr.
Liying Cheng. Her tremendous support led me through two degrees and exciting research
experiences. Further, I thank my committee member Dr. Don Klinger for his stimulating
challenges and inspiring discussions that helped to broaden my understanding of test
fairness. My appreciation also goes to Dr. Benjamin Kutsyuruba who provided insightful
comments and helpful suggestions. This dissertation would not have been possible
without their guidance and supervision.

I would like to give big thanks to my “counselors”, Dr. Nancy Hutchinson and Dr.
Han Han, who offered unfailing support during my rough patches along this educational
pursuit. I want to acknowledge Dr. Rebecca Luce-Kapler, Marlene Sayers, and Mary
Acreman, for their understanding and assistance at the hardest time of my learning. My
thankfulness goes to Dr. Anoop Gupta, Jane Rogers, Jean Marshall, Zack Morrison, and
my buddies Ying, Jia, Christine, Yi, King, Allison, Judy, and many more. I owe thanks to
Prof. He Lianzhen who helped with data access as well as my research participants, Liu
in particular, who shared their perceptions with me. Thanks also go to the Ontario
Graduate Scholarship program and Queen’s University for their financial support that
facilitated my academic studies. In addition, I want to thank my dissertation examining
committee—Dr. Janna Fox, Dr. Li-Jun Ji, and Dr. Nancy Hutchinson—for culminating
my studies in a rigorous academic defense that provoked my thinking.
Finally, I wish to thank the stakeholders in the overall picture of this journey—my parents who flew in to look after the kids when I needed them the most, my in-laws who provided wavering support taking care of the kids, and my sister and her family who helped to connect me with potential participants and drove me around for my research interviews. Thank you, most of all, to my husband Edmond, son Andy, and daughter Lauren for your patience and tolerance of my absent-mindedness and harshness when I had to meet all sorts of deadlines. After all, it is all about us.
Table of Contents

Abstract.................................................................................................................................................. ii
Co-Authorship........................................................................................................................................ iv
Acknowledgements............................................................................................................................... i
List of Figures .......................................................................................................................................... v
List of Tables ........................................................................................................................................ vi
CHAPTER ONE INTRODUCTION ........................................................................................................ 1
  Purpose .................................................................................................................................................. 4
  Research Context ................................................................................................................................. 5
  High-stakes Testing ............................................................................................................................... 11
  Test Fairness ......................................................................................................................................... 14
  Overview of the Manuscript-method Dissertation ............................................................................. 20
    Rationale of the Overall Research ..................................................................................................... 20
    Structure of the Dissertation ............................................................................................................. 22
CHAPTER TWO TEST FAIRNESS: PAST AND PRESENT ........................................................................... 26
  Early Development ............................................................................................................................... 27
    The Imperial Examinations and Other Similar Systems ..................................................................... 28
    Concerns of Bias ................................................................................................................................. 31
    The Psychometric Advances in Bias Detection .................................................................................. 34
  Empirical Studies on Test Fairness ....................................................................................................... 37
    Equality ............................................................................................................................................... 38
    Lack of Bias ........................................................................................................................................ 41
    Equity ................................................................................................................................................ 45
    Comparability across Different Forms .............................................................................................. 49
  Testing Standards, Guidelines, and Codes of Practices ........................................................................ 51
    Stakeholder Groups ............................................................................................................................ 54
    The Scope of Test Fairness ................................................................................................................ 55
      The Testing Process ......................................................................................................................... 56
      Opportunity to learn (OTL) ............................................................................................................... 59
      Impact of Test Use ........................................................................................................................... 60
    Fairness and Validity .......................................................................................................................... 61
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Conceptual Work</td>
<td>64</td>
</tr>
<tr>
<td>The Perspective Focusing on the Testing Process</td>
<td>65</td>
</tr>
<tr>
<td>The Perspective Focusing on the Socio-cultural Context</td>
<td>68</td>
</tr>
<tr>
<td>Summary of the Literature</td>
<td>72</td>
</tr>
<tr>
<td>Willingham’s Conceptual Framework and Operational Model</td>
<td>74</td>
</tr>
<tr>
<td>Willingham’s Conceptual Framework</td>
<td>74</td>
</tr>
<tr>
<td>Operational Model</td>
<td>78</td>
</tr>
<tr>
<td>CHAPTER THREE DIF INVESTIGATIONS ACROSS GROUPS OF GENDER AND ACADEMIC BACKGROUND IN A LARGE-SCALE HIGH-STAKES TEST</td>
<td>82</td>
</tr>
<tr>
<td>Introduction</td>
<td>82</td>
</tr>
<tr>
<td>Differential Item Functioning</td>
<td>85</td>
</tr>
<tr>
<td>Gender and Test Performance</td>
<td>88</td>
</tr>
<tr>
<td>Academic Background and Test Performance</td>
<td>90</td>
</tr>
<tr>
<td>Method</td>
<td>92</td>
</tr>
<tr>
<td>Subjects</td>
<td>92</td>
</tr>
<tr>
<td>The Graduate School Entrance English Examination</td>
<td>93</td>
</tr>
<tr>
<td>Data Analyses</td>
<td>95</td>
</tr>
<tr>
<td>SIBTEST</td>
<td>95</td>
</tr>
<tr>
<td>Content Analysis</td>
<td>96</td>
</tr>
<tr>
<td>Results</td>
<td>98</td>
</tr>
<tr>
<td>Descriptive Statistics and Test Evaluation</td>
<td>98</td>
</tr>
<tr>
<td>SIBTEST Results</td>
<td>99</td>
</tr>
<tr>
<td>Content Review</td>
<td>101</td>
</tr>
<tr>
<td>Gender DIF and DBF</td>
<td>101</td>
</tr>
<tr>
<td>Academic background DIF and DBF</td>
<td>103</td>
</tr>
<tr>
<td>Discussion</td>
<td>104</td>
</tr>
<tr>
<td>SIBTEST Findings</td>
<td>105</td>
</tr>
<tr>
<td>Content Analysis</td>
<td>106</td>
</tr>
<tr>
<td>Conclusion and Implications</td>
<td>109</td>
</tr>
</tbody>
</table>
CHAPTER FOUR FAIRNESS IN A LARGE-SCALE HIGH-STAKES LANGUAGE TEST: ADMINISTRATORS’, TEACHERS’, AND TEST TAKERS’ PERCEPTIONS

Introduction .......................................................................................................................... 112
The Graduate School Entrance English Examination ............................................................ 115
Conceptual Framework and Operational Model ................................................................. 118
Empirical Studies on Stakeholders’ Perceptions of Test Fairness ........................................ 123
Method ................................................................................................................................ 127
  Participants ....................................................................................................................... 127
  Instruments ...................................................................................................................... 130
  Data Collection and Analyses ......................................................................................... 131
Results ................................................................................................................................. 132
  Item Quality ..................................................................................................................... 134
  Standardized Administration ......................................................................................... 136
  Scoring ............................................................................................................................ 137
  Cut-scores and Admission Decisions ............................................................................. 140
  Overall Perceptions of the Fairness of the GSEEE ......................................................... 143
Discussion ............................................................................................................................ 145
  Knowledge of the GSEEE ............................................................................................... 146
  Perceptions of the Fairness of the GSEEE Policies and Practices ..................................... 147
  Item Quality ..................................................................................................................... 147
  Standardized Administration ......................................................................................... 149
  Scoring ............................................................................................................................ 149
  Cut-scores and Admission Decisions ............................................................................. 150
  Overall Perceptions of the Fairness of the GSEEE ......................................................... 152
Conclusion and Implications ............................................................................................... 154

CHAPTER FIVE TEST FAIRNESS OF THE GSEEE .................................................................... 157
Revisiting Willingham’s Conceptual Framework ..................................................................... 157
  Comparability .................................................................................................................. 157
  Process ............................................................................................................................ 160
  Participants ....................................................................................................................... 161
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation</td>
<td>163</td>
</tr>
<tr>
<td>The Tradition of Testing</td>
<td>164</td>
</tr>
<tr>
<td>The Testing-centred Meritocracy</td>
<td>165</td>
</tr>
<tr>
<td>Pursuit of Efficiency</td>
<td>167</td>
</tr>
<tr>
<td>Implications</td>
<td>170</td>
</tr>
<tr>
<td>Implications for Practice</td>
<td>170</td>
</tr>
<tr>
<td>Implications for Policy</td>
<td>172</td>
</tr>
<tr>
<td>Implications for Theory</td>
<td>173</td>
</tr>
<tr>
<td>Limitations and Future Research Directions</td>
<td>176</td>
</tr>
<tr>
<td>Conclusion</td>
<td>179</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>180</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>213</td>
</tr>
<tr>
<td>Appendix A Ethics Clearance for Study One</td>
<td>213</td>
</tr>
<tr>
<td>Appendix B Ethics Clearance for Study Two</td>
<td>214</td>
</tr>
<tr>
<td>Appendix C Letter of Information for Content Reviewers</td>
<td>215</td>
</tr>
<tr>
<td>Appendix D Consent Form</td>
<td>217</td>
</tr>
<tr>
<td>Appendix E Content Review Questionnaire</td>
<td>218</td>
</tr>
<tr>
<td>Appendix F Item Statistics Output</td>
<td>219</td>
</tr>
<tr>
<td>Appendix G Letter of Information for Test Takers</td>
<td>221</td>
</tr>
<tr>
<td>Appendix H Letter of Information for Program Administrators</td>
<td>223</td>
</tr>
<tr>
<td>Appendix I Letter of Information for English Teachers</td>
<td>224</td>
</tr>
<tr>
<td>Appendix J Consent Form for Test Takers</td>
<td>225</td>
</tr>
<tr>
<td>Appendix K Consent Form for Program Administrators</td>
<td>227</td>
</tr>
<tr>
<td>Appendix L Consent Form for English Teachers</td>
<td>228</td>
</tr>
<tr>
<td>Appendix M Recruitment Poster for Focus Group with Test takers</td>
<td>229</td>
</tr>
<tr>
<td>Appendix N Interview Protocol with Program Administrators</td>
<td>230</td>
</tr>
<tr>
<td>Appendix O Interview Protocol with English Teachers</td>
<td>231</td>
</tr>
<tr>
<td>Appendix P Interview Protocol for Focus Group with Test takers</td>
<td>232</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: The Graduate School Entrance Examination System................................. 9
Figure 2: An Operational Model for the GSEEE.................................................. 79 & 120
List of Tables

Table 1: A List of the Standards, Guidelines, and Codes of Practices ....................... 52-53
Table 2: Description of the GSEEE Administered in 2009 ......................................... 94
Table 3: Background Information of Content Reviewers ............................................. 97
Table 4: Results of Descriptive Statistics .................................................................... 99
Table 5: Results of SIBTEST Analysis ........................................................................ 100
Table 6: Results of Content Analysis ......................................................................... 101
Table 7: National Cut-Scores of the 2012 Administration ........................................... 117
Table 8: Background Information and Admission Status of the Test Takers ................. 129
Table 9: Themes and Code Frequencies ..................................................................... 133
CHAPTER ONE
INTRODUCTION

Large-scale high-stakes tests are often used as a means to classify, select, and judge individuals. They serve as “both door-openers and gatekeepers” (Bachman & Purpura, 2008, p. 456) to include individuals in, or exclude individuals from, academic communities and professional associations. For test takers, high-stakes testing is seen as a door-opener to a pathway that leads to educational resources, opportunities, and success. Policy makers, teachers, administrators, and academic association officials, on the other hand, tend to treat high-stakes testing as a gatekeeping device. For them, high-stakes testing is used to determine if test takers achieve a certain level of proficiency and competency, so these test takers are able to be engaged in academic communities successfully. The features of tests serving as both door-openers and gatekeepers highlight their important educational and social functions.

Concerns about test fairness among stakeholders are paramount in the milieu of high-stakes decision-making involving achievement, admission, and certification. Test fairness has been the subject of much discussion in educational research (Camilli, 2007), especially with the recent worldwide trend towards using testing and assessment for standards-based educational reform (Davis, 2009; Gipps & Stobart, 2009). Educators, students, parents, and the public voice their opinions on the fairness of testing through various channels, such as academic forums, newspaper editorials, and Internet discussions (e.g., fairtest.org). Test fairness has also been widely debated in court (Connor & Vargyas, 1992; Cumming, 2008; Mehrens & Popham, 1992). Many cases in relation to unfair treatment have been brought to courts and legislative bodies around the
world including Australia, the United States, and the United Kingdom (Buckendahl & Hunt, 2005; Cumming & Dickson, 2007). These legal cases related to large-scale high-stakes testing have challenged various aspects of testing policies and practices such as accommodations for students with special needs, the appropriateness of test use, and the availability of sufficient resources and learning opportunities (Cumming, 2008).

Fairness has been an important topic in China since the implementation of the Imperial Examinations (also called the Civil Service Examinations, 606 AD-1905 AD) when tests were used to select capable candidates for government positions across the whole country. Currently, high-stakes testing continues to play a key role in the Chinese educational system (Berry, 2011). Large-scale high-stakes entrance examinations are a major component in the selection of students who wish to attend “Key” secondary schools as well as undergraduate, master’s, and doctoral programs at universities and colleges. As with extraordinary economic transformation, educational development, and political reforms in China in recent years, test fairness has been a topic of much discussion (Liu, 2005; Xie & Li, 2000). Educational and economic development has greatly contributed to the need and affordability of individuals to obtain education and higher degrees in China. However, access to higher education in China remains very competitive, despite dramatic expansion in Chinese higher education and noticeable increases in enrollment in the late 1990s and early 2000s (Ding, 2007). There are continuous discussions over whether or not large-scale high-stakes testing provides fair opportunities to determine who will have access to higher education (Yang, 2007). Some of these discussions and disputes have been, as in the rest of the world, brought to courts for legal rulings (Exam8, 2010; Sohu China, 2006).
Given the significance of fairness in high-stakes testing, my dissertation research investigates fairness of one of the large-scale high-stakes tests in China—the Graduate School Entrance English Exam (GSEEE). The GSEEE is a national, standardized test that measures test takers’ knowledge of English and their abilities to use the language for either a degree program or research, and provides information for educational institutes to select candidates for their master’s programs (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. The GSEEE also influences other stakeholders such as graduate program administrators, teachers and supervisors of master’s students, parents of test takers, testing material publishers, test developers and item writers, educational policy makers, and future employers (He, 2010).

To clarify, fairness in this dissertation is used broadly to include policies, guidelines, practices, or perceptions with regards to test fairness. While polices provide general fairness principles, guidelines usually include the steps to be followed in practice as a consistent and repeating approach. Practices refer to more specific actions and activities actually conducted in testing, which, ideally, implement policies and guidelines and ensure fairness. Finally, perceptions on whether policies, guidelines, and practices are fair or not may vary from individual to individual, from group to group. I use specific terms, for example, perceptions of test fairness, where the discussion focuses on this specific aspect.
Purpose

This doctoral research has two central purposes. First, by using a psychometric technique to examine the presence of potentially biased items in one administration of the GSEEE, the research aims to provide empirical information regarding the fairness of test items from a psychometric perspective. Psychometrics is the field of quantitative study that examines cognitive and behavioral measurements (Jones & Thissen, 2007). Second, in addition to the focus on the potential bias in test items, I used a qualitative approach to examine stakeholders’ perceptions of the fairness of the GSEEE policies, guidelines, and practices across the testing process from test design and development, administration, scoring, to score-based test use. I also examined stakeholders’ general conclusions of the fairness of the GSEEE. Through discussions with stakeholders regarding their perceptions and general conclusions of the fairness of the GSEEE, the research has the potential to provide empirical evidence from the perspectives of stakeholders and advance our understandings regarding priorities and major aspects influencing stakeholders’ thoughts about the fairness of the GEEEE. Guiding my doctoral research are the following overarching questions:

(1) How do the GSEEE items bring potential bias, if any, toward test taker groups (e.g., groups based on gender and academic background)?

(2) How do the GSEEE test stakeholders (e.g., graduate program administrators, English teachers, and test takers) perceive the fairness of the GSEEE for groups and individuals of test takers?
Research Context

Education in China is generally characterized by the nationally-run system, despite some signs of decentralization practices (Hawkins, 2000). Within the Chinese educational system, examinations play a crucial role (Berry, 2011; Cheng, 2008). At school, students must undergo many levels of examinations as soon as they start their schooling in Grade 1, from small weekly quizzes to national, high-stakes testing. These national, high-stakes tests determine if students can enter “Key” secondary schools, colleges or universities, graduate schools, and professional organizations such as civil service and legislative communities. For students, doing well on these tests is the key to their academic success as well as the success in their life in general. A graduate degree such as a master’s degree remarkably increases life opportunities in China, where society, including the job market, is very competitive. The role of testing is intensified by the “Key School” mechanism, which has been in use since the 1950s (Berry, 2011; You, 2007). A small number of secondary schools and universities are selected as key schools and universities. Students compete fiercely for enrollment in these key schools and universities, which have excellent physical conditions, resources, and qualified teachers. Entrance examinations are used as a major tool to decide who can be admitted into these “Key” schools and universities.

On one hand, testing remains firmly in the Chinese educational system to serve educational and social functions. Frequently, students consider passing tests as their major learning goal; teachers follow test specifications to guide their teaching; and administrators use test scores to decide if students can be enrolled in their programs (Qu & Xiong, 2003). In China’s centralized, top-down political system, testing is also
believed to be an important tool to achieve social goals and distribute resources and services (Wang, 2010). On the other hand, more and more scholars and educators have used empirical data, rather than personal reflection, to support their arguments (Zhao, et al., 2008). Researchers have increasingly investigated if large-scale, high-stakes testing provides test takers with fair opportunities to demonstrate their ability in terms of test items (Liu & Wu, 2003), promotes non-learning activities and test-wiseness strategies (He, 2010), has an tendency to narrow curriculum and only focus on tested content (Qi, 2005), and brings enormous psychological pressure on students (Yu & Suen, 2005). The combination of a very long history of using tests in Chinese society and the relatively new development in conducting educational research and using empirical data to examine various issues related to testing, test quality in particular, has provided unique challenges, as well as opportunities, for fairness research.

The Graduate School Entrance English Exam (GSEEE) is one of the measures within the selection system of the Graduate School Entrance Examinations (GSEE) that is used to determine whether applicants can gain annual admission into a master’s program in China. After People’s Republic of China was established in 1949, the GSEE was briefly adopted in the 1950s and 60s followed by Culture Revolution (1966-1976), during a period when most educational activities were terminated. The GSEE, including the GSEEE, was restarted in 1978. There have been some major changes with the GSEE since its restart (Wang, 2009). The current GSEE includes the first round of preliminary examinations and the second round of re-examinations. While the first round of preliminary examinations is set and arranged in a unitary way by the Ministry of Education (MOE), the second round of testing is determined by individual institutes,
including important decisions such as time, place, content, and subjects. The first round of preliminary examination includes four tests: Foreign Languages (English, Japanese, or Russian), Political Science, and two additional subject areas. Each year, the National Education Examinations Authority (NEEA), appointed by the MOE, prepares test papers of Foreign Languages, Political Science, and some subject areas (e.g., Mathematics, Computer Science). A large majority of test takers choose English as their foreign language and write the GSEEE.

The first round of preliminary examinations plays an important role in admission decisions. Only those test takers whose scores surpass the cut-scores set by the MOE or select universities on all the four tests and total scores may enter the second round of re-examinations. Except for 34 universities, whose master’s programs have the flexibility to set up their own cut-scores, all the other institutes, over 900 of them, have to adopt national cut-scores. There are a number of different national cut-scores for each test based on ethnicity, areas, and disciplines. For example, in 2012, the whole country was divided into two broad region categories (Regions A and B) and 16 subcategories within each category based on ethnicity and disciplines (MOE, 2012). Cut-scores, including all the national and independently determined cut-scores, are decided based on a quota system—Enrollment Planning Objectives (EPOs), which is stipulated by the MOE and specific education committees at the provincial level (Wang, 2005). All the results from the first (including the GSEEE) and second rounds are valid for only one year during the application process. There are a large number of test takers every year and the competition is stiff. For example, the total number of GSEE test takers in 2011 reached approximately 1.51 million, and the acceptance rate was 32.75% (MOE, 2011).
The GSEE is under a three-level management system: the MOE and NEEA, the provincial examination board and admission offices, and the institutes that provide master’s programs (Chen, 2005; see Figure 1). Among the three levels, the MOE is mainly in charge of the following issues: (1) establishing the policies, guidelines, and methods for graduate school admission; (2) formulating the annual EPO in consultation with the relevant ministries; (3) prescribing the national test subjects; and (4) implementing the national tests. The National Education Examinations Authority (NEEA), appointed by the MOE, is responsible for test design, development, personnel training in item writing, and test evaluation of the national tests including the GSEEE (Liu, 2010). The second level, the provincial examination board and admission offices, is responsible for (1) the implementation of the graduate admission guidelines, policies, and methods that are advocated by the MOE, (2) test takers’ qualifications check and registrations, (3) test printing, distribution, and security, and (4) coordination, supervision, and organization of test administration, and scoring. At the third level, individual educational and research institutes take specific responsibilities that are closely relevant to their own admission needs. Their duties include (1) outlining the admission scheme, (2) test design and development of non-national subject tests in the first round, (3) selecting qualified candidates for sitting the second round of re-examinations, (4) test design, development, administration, and scoring of the second round of re-examinations, and (5) admitting test takers according to nationally-stipulated polices and principles.
Figure 1. The Graduate School Entrance Examination system
As stated earlier, the GSEEE has two purposes: to measure English proficiency of test takers and to provide information for educational and research institutes for selecting candidates for master’s programs (He, 2010). According to the GSEEE test specifications, the GSEEE examines test takers’ English language abilities, including knowledge of the language (grammatical competence, textual competence, and sociolinguistics competence) and language skills (NEEA, 2009; 2012). Currently, the GSEEE assesses only reading and writing. Speaking has never been included in the GSEEE. Listening was tested briefly in the GSEEE between 2002 and 2005 (He, 2010). Speaking and listening are tested in the second round of re-examinations for those test takers who surpassed the cut-scores in the first round of preliminary examinations, including the GSEEE.

The GSEEE has important consequences for its test stakeholders, such as test takers, graduate program administrators, teachers, test designers, policy makers, and testing material publishers. To facilitate this doctoral research, I classified these stakeholders into groups. In educational testing and assessment, test stakeholders are traditionally separated into three groups: test developers, test users, and test takers (AERA, APA, & NCME, 1999; ILTA, 2007). Within the context of the GSEEE, I conceptualized test developers as those who construct, administrate, score, and evaluate the GSEEE as well as those who set testing policies and guidelines regarding the GSEEE. Alternatively, test developers in this research mainly referred to those who serve on the MOE and NEEA and work for the MOE at the provincial level. Test users were those who interpret the GSEEE scores (having no right to select tests for decision-making in the first round within the context of the GSEEE) and make various decisions based on the
GSEEE, such as individual institutes and graduate program administrators, teachers, parents, and testing material publishers. Among test users, graduate program administrators and teachers are the two most salient groups. Although graduate program administrators and teachers have limited roles in deciding cut-scores (cut-scores are related to the national quota determined by the MOE), they are the most important test users since they make admission decisions and/or take actions based on the GSEEE test scores. Finally, test takers were those who take the GSEEE in order to obtain admission into a master’s program in China. Collecting empirical information regarding the fairness of the GSEEE, particularly from the perspective of test taker and user groups (Crocker, 2003), may be helpful to support the fairness claim of the GSEEE and enhance in-depth understandings of the concept of test fairness in China.

High-stakes Testing

Testing has a long history. It originates from the Chinese Imperial Examinations (also called the Civil Service Examinations) when ancient China selected the best potential candidates to serve as administrative officials in government. The Imperial Examinations are called the KeJu in Chinese (科举). Ke means “function” and Ju means “many persons gathering in one room and sharing” (Zhang & Zhong, 2003, p. 254). It is commonly regarded as starting around the year 606 and officially ending in 1905 (Yu & Suen, 2005). In order to be government officials, candidates were required to pass through three levels of tests from local, prefectural to palace testing over three years. Only candidates who passed the lower level of the examination were qualified to attend the next level. Those who failed had to wait for another three years before they could try again. The most successful individuals were identified to initially serve as scholars in the
imperial secretariat known as the Hanlin Academy (翰林院). From those positions, the
scholars might be promoted to serve as district magistrates, governors, national
departmental ministers, or even prime ministers and grand councillors (Yu & Suen,
2005). With such positions came all the legal privileges, power, reputation, and financial
rewards for the candidate and their entire extended family. The Imperial Examinations
were considered fairer than other selection practices that were based on patronage or birth
(Gipps & Stobart, 2009). Testing provided a level playing field, and test takers took the
same test under the same conditions regardless of their blood, family background, or
social class. The Imperial Examinations were the central focus of a national-orchestrated
selection system and test-driven education in Ancient China (Yu & Suen, 2005).

This examination system influenced many countries and it was later applied to
general education (Elman, 2013). Neighboring Asian countries such as Vietnam, Korea,
and Japan implemented similar civil service testing systems (Cheng, 2010; Gipps, 1999).
In Europe, school graduation exams were introduced for the standard school exit and
university entrance qualification in the 1800s (Broadfoot, 1979). High-stakes testing has
enjoyed a prominent role in deciding admission, achievement, aptitude, and certification
in various contexts in the 20th century (Gipps, 1999). Since the beginning of the 21st
century, more attention than ever before has been paid to issues such as accountability,
efficiency, and cost-benefit analysis of human resources in all walks of life. As a
consequence, testing has been extended to an increasing number of contexts worldwide
(Moses & Nanna, 2007). Currently, high-stakes testing is entrenched in the educational
landscape of countries such as China, Japan, Korea, India, Nepal, the United States, the
United Kingdom, Australia, France, Denmark, Sweden, Germany, Romania, and the
Czech Republic (Johnson, 2009; Minarechová, 2012; Phelps, 2005, 2008). On one hand, passing tests for individuals in these countries has important benefits, such as school and program admissions, diplomas, scholarships, certification and licensing to practice in a profession. On the other hand, failing has significant disadvantages, such as losing scholarships or further learning opportunities, being forced to take remedial classes until the test can be passed, or not being able to obtain educational opportunities, professional licenses, and find employment.

Despite an increasing number of contexts that rely on testing, high-stakes testing remains a controversial topic in education. Some researchers stress the important contributions of high-stakes testing with respect to representing effective methods for achievement, prediction, and selection (Jennings, 1998; Thurlow, 2002). Such a stance generally reflects the focus of competition, efforts, and individual responsibilities, highlights accomplishment or academic merit, stresses utility, and embraces economic productivity (Phelps, 2005; Walberg, 2004, 2011). However, other researchers criticize high-stakes testing and argue that it narrows the definition of intelligence and ability, has a limited connection with learning outcomes, undermines critical thinking and intellectual curiosity, and diminishes students’ self-esteem and motivation (Davis, 1998; Noddings, 2004). While some researchers highlight the limits of high-stakes testing (Johnson, 2008), some believe that high-stakes testing is not in principle unacceptable and that the problems generated can be controlled through better tests that more comprehensively represent the construct, use appropriate administrative and scoring policies, include multiple criteria for decision-making, and support better teaching practices (Elgin, 2004; Phelps, 2005).
In addition to the discussion of the negative impacts of testing on learning and teaching, some scholars question the merit-focused value of high-stakes testing. They point out the importance of equalizing educational opportunities and diversity of the student population (Flippo, 2003; Zwick, 2002). These scholars take into account the reality that different groups of test takers with different backgrounds often have unequal access to quality education (Moss et. al., 2005). They believe that test performance and decision-making that separates test takers from their learning opportunities and socio-economic status is illusory. Test fairness cannot be achieved if some test takers are deprived of learning opportunities that are available to others. Further, these scholars highlight the important goal of social diversity in student population in higher education (Spradlin, 2012). Such a goal cannot be achieved when proficiency and competency is mainly defined by test scores (Alon & Tienda, 2007). Overall, the tension of different value judgments between merit (defined by test scores) and accessibility/diversity is apparent in the discussion of high-stakes testing.

Test Fairness

Historically, fairness is an overriding topic of discussion with respect to high-stakes testing (Gipps & Stobart, 2009). It is often interwoven with another important concept in high-stakes testing—test validity (Xi, 2010). The relationship between fairness and validity has been defined differently (more discussion will follow in Chapter 2). In this dissertation, I take the stance that test fairness is highly intertwined with validity. Major issues discussed in test validity (Messick, 1989), such as construct underrepresentation, construct irrelevant variance, test takers’ or raters’ response processes, relations of test scores to other criteria, and test use, are also reflected in
various areas of fairness investigations. Such a stance is consistent with some of the literature in educational and language testing and assessment (Camilli, 2006; Willingham & Cole, 1997). This stance is also in alignment with the most recent edition of the Standards for educational and Psychological Testing\(^1\) that puts validity and fairness side by side as the foundations of testing (AERA, APA, & NCME, 2014).

While test fairness is an important research and practical issue, there is no widely accepted definition of test fairness that educational researchers, test developers, test users, and test takers all endorse. Test fairness has been conceptualized in many different, often conflicting, ways (e.g., AERA, APA, & NCME, 1999; Camilli, 2006; Cizek, 2009; Froese-Germain, 2001; Stobart, 2005). Some researchers discuss test fairness mainly from the statistical perspective, focusing on aggregating score-based evidence (Sackett, Borneman, & Connelly, 2008), some emphasize the social aspects of test fairness (Gipps & Stobart, 2009), and some argue that searching for test fairness is impossible and meaningless (Davies, 2010). For example, Helms defines test fairness as "removal from test scores of systematic variance attributable to experiences of racial or cultural socialization” (2006, p. 845). In contrast, Gipps and Stobart (2009) believe there is no cultural neutrality in assessment, and fair assessment practices need to take account of the social and cultural aspects. Halpern states that “fairness is both a statistical concept and a matter of societal values, and frequently these two types of fairness are at odds with each other” (2000, p. 58). Spann (2000) first contends fairness is “a philosophical idea which does not exist in the natural world,” but then states that fairness can be enhanced “through a synergistic, co-operative communication cycle among the test developers,\(^1\)

---

\(^1\) The most recent revision of the Standards will be published in July 2014; however, its table of contents has been printed for readers to browse in AERA 2014.
uses and examinees” (p. 37). Davies (2010) asserts that the pursuit of fairness is in vain “first because it is unattainable and second because it is unnecessary” (p. 171). These conceptualizations take distinctive stances and evince sharp differences as well as much overlap. In the following section, I briefly review the historical development of test fairness to provide context and insight into the logic and legitimacy of investigating fairness in this dissertation. More detailed, historical information regarding test fairness will be provided in Chapter 2.

The discussion related to test fairness has a very long history, with evolving foci and enlarging scope. As introduced earlier, high-stakes testing stems from the Chinese Imperial Examinations, which promoted equal treatment through various testing practices such as double marking. Selections were considered fair and just because test takers were treated identically (at least roughly) with the same testing condition, using the same scoring procedure. Those who performed well demonstrated their ability and attributes; hence they were selected to serve in government (Gipps, 1999). This practice focusing on equality prevailed in China and around the world until the advent of public education during the late 19th and early 20th centuries in Europe and North America. Because of the need to identify students capable of higher education and professional associations, intelligence testing (IQ) was adopted to differentiate those with potential and without (Hanson, 1993). However, some educators found persistent racial differences in IQ scores and raised concerns of test fairness such as potentially biased content and differences in the educational opportunities for different test taker groups (Cronbach, 1975; Du Bois, 1918).
In the late 1960s and early 1970s, the Civil Rights Movement in the United States gave social identity to the topic of test fairness. Given noticeable group differences in score-based decisions, particularly related to Black and Hispanic students, various statistical models for fair decision-making were put forward in the measurement community (Cleary, 1986; Cole, 1973; Darlington, 1971). Nevertheless, researchers did not agree on a preferred model to ensure fairness in decision-making. Later in the mid-to late 1980s, research emerged on a broad scale in terms of how test items functioned differentially, and might bring advantages or disadvantages toward different groups of test takers (Holland, 1985; Holland & Thayer, 1988). Various techniques of Differential Item Functioning (DIF) were tested and used to judge whether test items functioned in the same manner for different groups of test takers. However, the determination of item bias was often inconclusive and it was often not clear which aspect of a test item caused DIF (Walker, 2011).

While test fairness continues to be associated with bias and the measurement aspect of testing (Helms, 2006; Sackett, Borneman, & Connelly, 2008), the discussion about test fairness now attaches more importance to value judgments related to the testing process as well as the influence of broad social, cultural, and political considerations (Cumming, 2008; Gipps & Stobart, 2009). Although the current literature, mostly conducted in Western testing contexts, generally recommends that fairness requires that groups and individuals of test takers take the same or equivalent tests, under the same or accommodated conditions, and their performances are evaluated using the same scoring schemes, it is unclear whether these practices are accepted and adopted unanimously all over the world (e.g., test accommodation practices). Value judgments are particularly
evident in score-based decision-making. While some literature highlights equal treatment of test takers in decision-making (AERA, APA, & NCME, 1999), others encourage the practice of setting different cut-scores for different test taker groups and state that decision-making should consider test takers’ backgrounds and unbalanced learning opportunities and ensure equal representation in student bodies (JAC, 1993; Zwick, 2004). Furthermore, some researchers contend that it is important to shift toward the analysis of the socio-cultural context in conceptualizing test fairness (Moss et. al., 2005). Such views focus on reasoning of various social, cultural, educational, political, economic, and philosophical elements (Camilli, 2013; Kunnan, 2008). Despite various foci and scopes, more and more researchers have discussed test fairness from a broad point of view beyond psychometric considerations (Davis, 2009).

Here I offer a brief discussion of five revisions of the Standards for Educational and Psychological Tests, the preeminent resource regarding educational testing in North America, to explain how the thinking related to fairness has evolved. Such evolution is demonstrated in two areas: the importance and articulation of test fairness as well as the relationship between bias and fairness. While the 1954 and 1966 editions did not mention fairness, the 1974 edition states that "some unfairness may be built into a test" (p. 2). Fairness in the 1974 revision mostly refers to predictive bias in test use, which is consistent with the technical advances during this period of time. Alongside the advancements in DIF statistical techniques, the 1985 edition of the Standards discusses item bias and predictive bias. It also states that "fairness is not a technical psychometric term; it is subject to different definitions in different social and political circumstances" (p. 13). However, the 1985 Standards does not explain how fairness differs from bias,
what and why test fairness involves in the North American context, and in what way the measurement community could help to address fairness concerns. In the most recent revision of the Standards in 1999, Part 2, which includes four chapters, is completely devoted to test fairness. The gathering of the four chapters under one heading reflects the growing recognition of the importance of test fairness. In the 1999 version, four interpretations are included in defining test fairness: lack of bias, equitable treatment in the testing process, equality of test outcomes, and opportunity to learn materials in achievement testing (more discussions will follow in Chapter 2). Bias in the 1999 version is defined, clearly, as a technical term, referring to “construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees” (1999, p. 76). This is the first time that the Standards articulates how fairness differs from bias. However, as shown in the section Can consensus be achieved?, the 1999 version takes a stance that test fairness is “subject to different definitions and interpretations in different social and political circumstances” (1999, p. 80). Even for the four interpretations of test fairness outlined in the 1999 version, there remain different views. As commented by one of the contributors to the 1999 Standards (Green, 1998), the first two interpretations have generally been accepted, while the last two are far less straightforward and remain controversial in educational testing and assessment. Overall, the four interpretations indicate the importance of the concept of test fairness and the evaluation of the definitions regarding test fairness, which is currently perceived as larger than, and inclusive of, bias.

In summary, the evolution of the thinking related to test fairness reflects the epistemological underpinnings that now recognize the multi-dimensionality of test
fairness, the role of subjective value judgments within the testing process (e.g., decision-making), and the influence of broad social, cultural, and political considerations.

Historical and ongoing changes in our conceptions of test fairness demonstrate that fairness is a continuously evolving process. Through revisions in relevant documents such as the Standards and as the focus about test fairness evolves, it has become evident that fairness is a complex issue and an enduring concern in education. More empirical research is needed to examine this complex concept, especially within the Chinese context. Although the ideal of test fairness has been pursued since the Imperial Examinations, limited research has been conducted and it is unclear whether the concept of test fairness has evolved in China.

Overview of the Manuscript-method Dissertation

This doctoral research uses a manuscript approach, which is different from the traditional format of the doctoral dissertation (Queen’s University, 2013). This dissertation consists of a comprehensive review of previous research, two separate manuscripts, and a final, integrative discussion chapter. The first manuscript examines potential test bias and the second investigates stakeholders’ perceptions of the fairness of the GSEE. In the following section, I first provide an overall rationale for why the two studies are conducted in such a manner. I then describe the organization of the rest of the chapters as well as each of the manuscripts.

Rationale of the Overall Research

There are three explicit reasons for the selected research focus and design. First, in testing and measurement, there are concerted efforts to ensure that tests and test items are designed to be fair for different groups of test takers in test design and development.
One of the key, traditional aspects in obtaining this type of score-driven evidence is to detect bias in favour of, or against test takers from certain groups (e.g., gender, linguistic, or socio-economic status) that result in construct irrelevant differences in test scores (Cole & Zieky, 2001; McNamara & Roever, 2006). Differential item functioning (DIF) is one of the most commonly used methods to detect potentially biased test items. It is a statistical procedure for judging if test items are functioning in the same manner for different groups of test takers who are matched on proficiency level. As such, in Study 1, I examined if potentially biased items exist in one administration of the GSEEE. Using the traditional, psychometric approach along with content analysis, I investigated potential bias with groups based on gender and academic background, given China’s demographic information and the features of the GSEEE.

Second, in addition to potential bias in test design and development, the literature has discussed the extensive scope of test fairness and value judgments in deciding test fairness. Although the literature has identified some common features related to test fairness including lack of bias in test design and development (AERA, APA, & NCME, 1999; McNamara & Roever, 2006), it is unclear whether these features are accepted and adopted unanimously (e.g., test accommodation practices). Moreover, some aspects remain controversial and value judgments are particularly evident in what constitutes fairness in score-based decision-making (Camilli, 2013). Accordingly, in Study 2, my attention shifted from the potential bias in test design and development to a larger scope of fairness investigations from test design and development, administration, scoring, to score-based use. Study 2 may help us to identify key features of the fairness of the GSEEE as well as the fundamental reasons for such determinations.
Third, the literature has identified the importance of examining test fairness from a broad point of view beyond psychometric considerations such as DIF. Researchers in educational and language testing and assessment have pointed out the usefulness of the stakeholder approach collecting information based on their perceptions and reflections to provide evidence regarding the claims that a test makes (Crocker, 2003; Hamp-Lyons, 2000). There is a need, in particular, for stakeholders from outside the technical community (e.g., test users and test takers) to address issues of test fairness because their voices and perceptions tend to be under-researched. Hence in Study 2, I examined perceptions as expressed by the GSEE test users and test takers. Using an alternative lens by exploring test users’ and takers’ perceptions, Study 2 may provide empirical information regarding the fairness of the GSEE and their value judgments. Involvement of test users and test takers may help to identify problems related to the GSEE, which can be under-identified by traditional psychometric approaches such as DIF.

In summary, Studies 1 and 2 will offer empirical information regarding the fairness of the GSEE from different perspectives. There seems to be no research like this kind that has been conducted in language testing and assessment. This research has the potential to fill a void to examine test fairness from both psychometric and stakeholder perspectives and investigate the conceptualization of the fairness within a specific context—the GSEE.

Structure of the Dissertation

The dissertation is written in five chapters. In this Chapter 1, I have introduced the purpose of this research by articulating two general research questions. I have also described the research context and briefly discussed high-stakes testing and test fairness.
to provide foundational information. Finally, I have provided a rationale for a manuscript-method dissertation.

In Chapter 2, I will review four sources of literature related to test fairness. I begin with an introduction of the early development in test fairness. Then I review empirical studies that examined test fairness from various dimensions. After that, I evaluate and synthesize two types of texts: testing standards, guidelines, and codes of practices drafted by groups of authors as well as the fairness frameworks developed by individual scholars. The comprehensive review of the early development, empirical studies, and conceptual work will provide me with the solid foundation to identify a conceptual framework to guide this dissertation. In the end, an operational model will be developed to provide specific directions for fairness investigations in action.

The remainder of this dissertation is organized into two manuscripts, Chapters 3 and 4, followed by a general discussion of the findings, Chapter 5. The two manuscripts, Chapters 3 and 4, adopt a standard research procedure for journal publication. The two manuscripts consist of an introduction, literature review, method (subjects/participants, instrument, data collection and analysis), results, discussion, and implications. The slight difference between the two manuscripts and a regular journal paper is that the references for each manuscript are compiled at the end of the dissertation, along with references for other chapters, based on the dissertation requirement of Queen’s University. Due to the nature of the manuscript method, Chapters 3 and 4 are written as stand-alone documents for journal publication. As such, the same relevant information such as research context and key concepts will be repeated in the two manuscripts.
In Chapter 3, I first review the concept of DIF as well as the previous empirical studies using DIF. In line with the fairness conceptual framework outlined at the end of Chapter 2, I examine one traditionally researched aspect — potentially biased items towards different test taker groups in one GSEEE administration, using DIF and content analysis. Considering the demographic information of China and characteristics of the GSEEE test takers, I am particularly interested in two grouping variables: gender (female or male) and academic background (Humanities/Social Sciences or Sciences). As DIF detection has not been conducted in the GSEEE, this study will provide empirical information regarding if different groups of test takers are provided with fair opportunities to demonstrate their knowledge and skills. Results of this study may be used to reinforce current test design and development practices or may lead to modifications in test development practices.

In Chapter 4, I examine stakeholders’ perceptions of the fairness of the GSEEE across the testing process from test design and development, administration, scoring, to score-based test use, guided by the conceptual framework identified in Chapter 2. Among the GSEEE test stakeholders, test takers and two subgroups of test users (graduate program administrators and English teachers) are the most important groups; hence, I elicit information from these groups. Through one-on-one interviews with test users and focus group interviews with test takers, I explore perceptions of these stakeholders on the fairness of the GSEEE testing policies, guidelines, and practices as well as overall perceptions on the fairness of the GSEEE. The study will make contributions in obtaining empirical accounts of the fairness of the GSEEE from these stakeholders’ perspectives.
This study will also identify the major elements and priorities in considering the fairness of the GSEEE among the participants.

In the last chapter, Chapter 5, I provide a general discussion by tying the two studies together. Findings are considered in light of the fairness conceptual framework that is used to guide the research as well as the major conclusions concerning the fairness in the GSEEE. In Chapter 5, I also discuss general implications, acknowledge limitations, and suggest avenues for future fairness research.
CHAPTER TWO

TEST FAIRNESS: PAST AND PRESENT

Inquiry into fairness of a test has been recognized as a meaningful research activity to direct efforts to reduce or eliminate bias and discrimination against certain groups of test takers, create equal opportunities for test takers to demonstrate their knowledge and skills, and promote social justice (Gipps & Stobart, 2009; McNamara & Ryan, 2011). Although test fairness has been discussed widely in educational research as well as in public forums and court, one of the major difficulties in fairness research is how to define test fairness and to identify key issues for fairness investigations (Xi, 2010). Test fairness has been conceptualized in many different, often conflicting, ways (e.g., AERA, APA, & NCME, 1999; Camilli, 2006; Cizek, 2009; Cole & Zieky, 2001; Froese-Germain, 2001; Helms, 2006; Moss et al., 2005). There is much discussion in terms of the dimensions and scope of test fairness as well as how educational researchers respond to fairness inquiries methodologically. The challenge in fairness research lies in finding a framework that is theoretically sound, methodologically applicable, and empirically useful (Xi, 2010).

In an effort to elicit key issues related to test fairness to guide this doctoral dissertation, I draw on the conceptual and empirical literature in this chapter and identify a fairness conceptual framework that is appropriate for the research context. I review four sources of relevant literature: (1) the early development of test fairness, (2) empirical studies, (3) standards, guidelines, and codes of practices, and (4) individual conceptual work. First, I examine the early development of test fairness to provide a historical basis for the discussion of the concept of test fairness. Then, I review empirical studies that
investigated various dimensions of test fairness in different contexts. The empirical studies may also provide insights into research methodologies. After that, I examine two classes of more recent literature in the conceptualization of test fairness: testing standards, guidelines, and codes of practices adopted by educational and language testing and assessment organizations, and the conceptual frameworks written by individual scholars. Standards, guidelines, and codes of practices are developed by large committees to provide general agreement within or across organizations and guide step-by-step testing practices. In comparison, individual conceptual work, which can inform standards, guidelines, and codes of practices, usually offers original arguments and furthers discussions in the academic community. Despite significant differences in intent and construction, both classes of literature provide important information concerning the dimensions, scope, and complexity of fairness investigations. Through the syntheses of the early development and empirical studies as well as standards, guidelines, and codes of practices, and individual conceptual work related to test fairness, I intend to identify a conceptual framework and propose an operational model to guide the fairness investigations in my research.

Early Development

My reading has enabled me to identify three distinct stages within the early development of the concept of test fairness: (1) the Imperial Examinations and other similar systems; (2) concerns of bias on IQ tests during the first half of the 20th century; and (3) the psychometric advances in bias detection during the second half of the 20th century. The early development provides information on the changing foci with regard to test fairness: from equality to bias, from promoting individual fairness to group fairness,
and from highlighting general fairness procedures to evidence-based research and psychometric advances.

*The Imperial Examinations and Other Similar Systems*

As briefly introduced in Chapter 1, interest in test fairness can be traced to the Chinese Imperial Examinations (also called the Civil Service Examination, 606 AD-1905 AD). Similar to any other civilizations, ancient China had multi-layered social groups and classes, such as emperors, gentry, peasants, artisans, and merchants. The selection for civil service positions in early ages was generally based on family background and connections (Ebrey, 2010). In the Han Dynasty (206 BC-220 AD), the emperor started an early form of the Imperial Examinations: a group of outstanding young men who were recommended from different provinces by officials were educated by scholars with national reputations. After one year, these young men were tested, and those who succeeded in testing were given official positions (Franke, 1960). Subsequently, in 606 of the Sui Dynasty (581AD-618 AD), the system of the Imperial Examinations was officially established. The emperor decreed that exams would be the official criterion for those who were interested in government positions. This national policy allowed people from the lower social classes (e.g., poor peasants, artisans, and merchants) to have opportunities to take the tests, and potentially obtain civil service positions. In the Song Dynasty (907 AD-1276 AD), more and more candidates attempted the Imperial Examinations. The number increased from less than 30,000 in the early 11th century to about 400,000 test takers when the Song dynasty ended (Ebrey, 2010). By the Ming Dynasty (1368 AD-1644 AD), the examination system, consisting of three major progressive levels (local, prefecture, and palace) over three years, was fully developed.
The examinations consisted of “Five Studies” (military strategy, civil law, revenue and taxation, agriculture, and geography) as well as the “Four Books” and “Five Classics”, a set of philosophical works written by Confucius and his disciples. Almost anyone who wished to become a government official had to prove his merit and worth by passing the Imperial Examinations (Miyazaki, 1981). The Imperial Examinations, at least seemingly, provided a level playing field open for any candidates (commonly male), no matter whether they were peasants or merchants, to demonstrate their talent and compete for government positions.

The Imperial Examinations promoted the ideal of equality through equal, impartial treatment in testing across the whole country, regardless of people’s social classes and family background (He, 2012). During the three levels of testing, extremely rigorous procedures were used to provide test takers with fair and equal treatment, including double marking, examination avoidance of a conflict of interest, candidates being locked in a separate examination cell for three days, and test scripts with test takers’ names covered and being copied so that their handwriting would not be recognised (Cheng, 2010; Yu & Suen, 2005). The most successful candidates were summoned to the imperial palace for the final examination under the personal supervision of the Chinese emperors. The emperors would interview each candidate to determine the candidates’ knowledge and skills. Although recommendations based on family connections for government positions still existed, Chinese history is replete with written records of talented men from humble backgrounds (e.g., poor peasants and merchants) who succeeded on the Imperial Examinations, and ended up attaining high office such as district magistrates, governors, and prime ministers (Ebrey, 2010). The Imperial
Examinations lasted until the end of last dynasty of China—Qing Dynasty (1664 AD-1911AD).

Historically, the examination system helped the Chinese government choose individuals with the skills and talents required for governmental positions, promoted fairness through equal treatment of test takers, and prepared a large number of professionals for its pre-modern civilization (Hu & Seifman, 1987; Rosen, 1997). The Imperial Examinations played an important role in education and social mobility. The system ruled almost every aspect of education through tested knowledge (e.g., Confucianism), format (e.g., eight-legged essay\(^2\) 八股文), and skills (e.g., essay writing). Compared with the aristocratic system in which selection was grounded on group memberships such as hereditary title or social status, the Imperial Examinations were considered superior and fair because selection was based on merit and test performance under equal conditions. The examinations were viewed as a meritocratic solution to the widespread social problems of bribery, nepotism, and corruption in traditional China (Elman, 1991; Yu & Suen, 2005). Of course, an apparent challenge related to the Imperial Examinations was equal access to education and testing. Test takers were generally limited to males whose families could afford education and years-long examinations.

The Imperial Examinations influenced many other countries and regions including Korea, Japan, Vietnam, and Europe (Eckstein & Noah, 1993; Gipps, 1999).

\(^2\) The eight-legged essay (baguwen) is a written form of argument in the Imperial Examination in China. The essay has to be composed according to a strict pattern called the “eight-legged essay”, with an introduction, exposition, argumentation, and conclusion, both in two sections. Each “leg” has to be written in words that parallel its counterpart in the corresponding section. Even the number of characters or words to be used is regulated. This format later becomes synonyms of pedantry or triteness.
Ancient Vietnam implemented similar testing systems in the 11th century; the Jesuits introduced competitive examinations into their schools in the 17th century; and examinations were developed in northern Europe in the late 18th century (Gipps, 1999). During the 19th century, the Industrial Revolution marked a major turning point in history, and capitalist economies needed more skilled workers and professionals. These workers and professionals, who represented the rise of the middle class, realized that education was a means for upward mobility and testing provided a channel for them to have access to education, training, and opportunities. The growing middle class resulted in a need to abandon the previous practice by which selection was based on group membership of blood, patronage, and family history. The academic and professional communities started the new practice of using exams to select those who would receive education and professional membership. These communities set up examination boards and introduced exams to decide who were deemed suitable for training and education. For example, Britain introduced qualifying exams for the medical profession in 1815, and Oxford and Cambridge University set up examination boards and introduced entrance examinations in the 1850s (Gipps, 1999). Unfortunately, the European testing systems, similar to the Chinese Imperial Examinations, could not eliminate the advantages afforded by gender, social economic status, and wealth (Gipps & Stobart, 2009). Testing was generally restricted to males from relatively high socio-economic family backgrounds.

**Concerns of Bias**

During the late 19th and early 20th centuries, the booming industrial economy brought the advent of public education into Europe and North America. Educational
development in public education welcomed an unprecedented number of students. More and more children were enrolled in school. This situation resulted in a need to differentiate those capable for secondary and higher education from those for labour work (Gipps, 1997). Although originally developed to identify mental retardation, Intelligence testing (IQ) became a popular method to differentiate cognitive ability and decide learning opportunities amongst children (Hanson, 1993). The first IQ test that Alfred Binet developed in 1905 intended to examine verbal abilities to identify mental problems and provide these children with special help in coping with the school curriculum. Later, in 1916, the American psychologist Lewis Terman released a revised test—the Stanford-Binet test. This revised Standford-Binet scale aimed at differentiating children’s cognitive skills, and it was used as a defensible method to identify those capable for further education (Hanson, 1993). The Stanford-Binet test was used not only in general education for selection purposes, but also other purposes such as army recruitment and employment (Black, 2001).

Despite the wide use of IQ tests, some educators and researchers found there existed constant performance differences for different racial groups in the 1920s. They suspected that performance differences were derived from test content bias and different educational opportunities (Du Bois, 1920). They argued that IQ test items might be culturally biased in favour of test takers from the dominant culture, normally those white males from the middle class. These tests were subsequently found to reflect the values, cultures, and experience of item designers who were mostly white, male psychologists (Cronbach, 1975; Jenson, 1980). It was also argued that the socio-economic background of test takers could affect test performance, and IQ tests obscured the perpetuation of
unfairness and social inequalities in learning opportunities. Questions were raised about the use of testing to inherently discriminate against some groups from low socio-economic backgrounds. Overall, there were developing thoughts and views that the use of these tests, which were seen originally as fair tools, may not provide the fairness as was claimed.

Along with the implementation of these IQ tests was the development of psychometrics, which focused on the construction and validation of cognitive measurement instruments and educational tests (Brown & Thomson, 1921; Kuder & Richardson, 1937). Researchers emphasized fairness in developing educational tests. For example, Thorndike advocated the importance of methods “designed to free measurements from certain pernicious disturbing factors, notably unfair preparation for the test, inequalities in interest and efforts, and inequalities in understanding what the task is” (Thorndike, 1918, p. 23). With the infrastructure provided by emerging psychometric organizations such the Psychometrical Society founded in 1935 and the journal Educational and Psychological Measurement in 1941, important advances had been made in psychometrics, which motivated data collection in empirical studies. Quantitative methods and research started to gain prominence in educational research in the United States and Europe during this period of time (Jones & Thissen, 2007). The focus of using psychometric methods to investigate tests has strongly influenced the educational community through the second half of the last century and into the 21st century. This focus highlights the importance of using psychometric approaches to examine test fairness in my doctoral research.
The Psychometric Advances in Bias Detection

Major psychometric advances related to test fairness emerged in the late 60’s and early 70’s as the result of the Civil Rights Movement in the United States. During this period of time, cultural identity received growing attention. The public asked whether score-based decisions inherently discriminated against some groups, in particular, Black and Hispanic test takers. In response to the growing "public hostility" to testing (Messick, 1965, p. 136), the notion of culture-fair tests evolved, and there was a concerted effort to investigate the fair use of test results in the selection of candidates (Cole & Zieky, 2001). There was an intense burst of interest in test fairness research focusing on fair selection models, and various statistical models for fair decision-making were put forward (Cleary, 1968; Cole, 1973; Darlington, 1971).

Despite much attention and discussion on how to achieve fairness in decision-making, researchers and educators did not agree on a preferred model. For example, the Cleary Model argued that fairness would be achieved through selecting individuals who were most likely to succeed, regardless of their group membership (Cleary, 1968). Therefore, decisions based on the common regression line for each group and each individual provided the fairest decisions. In contrast, the Cole model required an adjustment of accept-reject rates for different groups through the use of different group-based regression lines, so that these groups would have the equal proportions of successful test takers to be represented in the student population. While the Cleary model defined equality in decision-making as fairness (similar to what the Imperial Examinations advocated), the Cole model demonstrated that test fairness needed to consider the diversity of the student population and take into account the reality that
different groups of test takers would have unequal access to quality education. This creates a dilemma. As pointed out by Willingham and Cole (1997), “the method [the Clearly Model] that seems fair to individuals is not fair to groups and the method [the Cole Model] that seems fair to groups is not fair to individuals” (p. 314).

Although there was no preferred model, researchers and educators agreed that fair selection was closely related to subjective value judgments. In a special issue of the Journal of Educational Measurement (1976) published almost four decades ago, the researchers concluded that issues related to fair selection may not be solved by psychometric models (Cronbach, 1976; Linn, 1976). Petersen and Novick (1976) pointed out that the definitions of fairness in decision-making depended on value judgments, and it was important to make the values explicit instead of hiding the values. The dilemma regarding which model provides fairness in selection reflects a contradiction between different values – focusing on the interests of individuals or groups. These different values, which permeate high-stakes decision-making, have been widely discussed in educational assessment, and questioned many times in the courts of different countries (Cumming, 2008; Smith, 2011; Zwick, 2006). The admission decisions and methods that are adopted to enroll subgroups of students from disadvantaged backgrounds continue to be a challenge in testing contexts (Brelend et al., 2002).

In addition to the discussions of fair selection models, the psychometric advances related to fairness expanded to another area – item bias. During the mid-to late 1980s, research emerged on the analysis of item bias in the measurement community. The research, commonly referred to today as “Differential Item Functioning” analysis, was introduced in a series of papers by Paul Holland and his colleagues at the Educational
Differential Item Functioning (DIF) is a statistical procedure for judging whether test items are functioning in the same manner for different groups of test takers. According to Roussos and Stout (2004), 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” (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 might occur if the existence of DIF is due to the situation in which test items measure nuisance dimensions that are not relevant to the underlying ability of interest. In the more than 30 years since Holland's series, the DIF framework has produced an enormous wealth of research and dozens of DIF procedures have been described in the literature (see a review by Clauser & Mazor, 1998).

Regardless, the literature agrees that DIF and bias are different — DIF is a necessary but not sufficient condition for bias (McNamara & Rover, 2006). There might exist legitimate reasons for differential functioning. What is concluded as potential bias or nuisance dimensions may depend on subjective judgments, most often, the review of disciplinary experts. Overall, this type of research, again related to value judgments, can be seen as an extension of the previous question with regards to fairness in item design of IQ tests in the first half of the 20th century. It is motivated by the desire to detect bias and ensure fairness for different groups of test takers in test design and development from the psychometric perspective (McNamara & Rover, 2006).

To conclude, this section reviewed the early development related to test fairness. Fairness was promoted and pursued through equal treatment among all individuals in the
Imperial Examinations and other similar systems, which enjoyed appeal and popularity in China and other parts of the world till the end of the 19th century. However, since the first half of the 20th century, the focus of test fairness went beyond equality among all test takers. Researchers and educators in the West had increasingly found ostensible group performance differences, racial differences in particular, in IQ tests, and they raised fairness concerns for these test taker groups. Attention had been shifted from equal treatment for all individuals to potential bias toward certain groups in decision-making and item design in the second half of the 20th century. Test fairness was motivated by the quest of reducing bias for test taker groups, in addition to equality for all individuals. Furthermore, fairness research in the second half of the 20th century used quantitative methods and statistical models to identify bias. Despite much progress using statistical models and measurement theories, value judgments permeated the discussions about which model should be used for decision-making as well as what was considered potential bias (MaNamara & Rover, 2006).

Empirical Studies on Test Fairness

Empirical studies related to test fairness have been conducted with various tests in various testing contexts. These empirical studies have typically focused on investigating one aspect at one time. For example, studies have examined differences in test performance due to differential item functioning (DIF) across various groups (Kunnan, 2000; Ferne & Rupp, 2007), construct-irrelevant test taker characteristics (Clapham, 1998), different item formats and forms (Bolger & Kellaghan, 1990; Taylor et al., 1998), accommodation practices (Finch, Barton, & Meyer, 2009), rater behaviors and reactions (May, 2009), and learning opportunities (Walpole et al., 2005). There have also been
studies that examine construct-irrelevant variance in test administration (Huang & Garner, 2009), predictive bias across various groups (Leonard & Jiang, 1999), the impact of various factors in decision-making (Breland et al., 1995), and stakeholders’ perceptions of the fairness of the test and test items (Fox & Cheng, 2005). These empirical studies offer insightful information concerning various dimensions of test fairness. In the following, I elaborate on empirical studies focusing on four main dimensions of test fairness:

(1) Equality (e.g., construct-irrelevant variance in standardized test administration and the scoring procedure);

(2) Lack of bias (e.g., test bias, predictive bias, stakeholders’ perceptions of the test and test items);

(3) Equity (e.g. the influence of accommodation practices on test performance);

(4) Comparability across different forms (e.g., the influence of parallel and modified formats and forms on test performance).

Equality

The dimension of equality has been a traditional focus of test fairness since the Chinese Imperial Examinations. Equality indicates sameness, meaning that all test takers are equal and they are treated as identically as possible, regardless of their group membership such as gender, race, language background, or other ascribed characteristics (Caldwell, Shapiro, & Gross, 2007). Generally, the relevant empirical studies focus on three areas of fairness investigations: standardized administration, impartial scoring, and equal opportunity to learn (OTL).
First, empirical studies generally agree on the importance of equality in standardized administration to ensure test fairness (Bridgeman & Schmitt, 1997; McCabe et al., 2001). In any test, test administration is the most public and visible aspect (Haladyna & Downing, 2006). Standardized administration occurs when a test is administered under uniform physical conditions and consistent administrations unless there is a good reason to adjust these procedures (Haladyna & Downing, 2006). The major function of standardization is to provide equal, consistent, and secure administration so each regular test taker can perform at their best (McCabe et al., 2001). Standardized administration is intended to ensure equal treatment of the test takers, which allows score comparisons among all regular test takers (those without special needs). Administration procedures also protect the security of the test to help to maintain the meaning and integrity of each test score. Cheating often poses a threat to equal treatment and jeopardizes test security and integrity of test scores. While the major reasons for standardized procedures are not to prevent cheating, cheating can be an issue when these standardized tests have important consequences for test takers. As an example, cheating in large-scale high-scales testing is a major concern in China, as described in academic research and newspapers (Marcus, 2013). Huang and Garner (2009) interviewed 17 test takers and 3 test administrators for a large-scale high-stakes language proficiency test in China. The results identified various issues related to test security. Cheating was perceived to be widespread during test administration of this national English test. Although a variety of measures had been used to ensure test security and prevent cheating, including moral education, punishment, security procedures, and detection
through analyzing testing responses, adverse impact was manifested and the goal of test fairness was not achieved.

Second, empirical studies highlight fair treatment in scoring, which requires impartiality without favoritism for any individuals (Willingham & Cole, 1997). Empirical research related to fair scoring has been undertaken to examine a variety of factors such as test form scoring comparability, scoring errors, rater favouritism and severity, and prompt choices (O’Loughlin, 2002; Schaefer, 2008). These factors may result in unfairness for test takers. For example, Lamprianou (2008) investigated the effect of test scores in a language exam when raters differed significantly in severity and self-selected questions (prompt choices) differed significantly in difficulty levels. Using the IRT Rasch model, the study concluded that test takers could benefit significantly from being marked by lenient raters and by responding to less demanding essay questions, which challenged the ideal of fairness in language testing. May (2009) examined the reactions of four raters to paired test takers, 12 in total, who had different levels of oral competency in a language communicative task. Results found that the key feature of the successful interaction perceived by the raters was two-way communications and group work. The individuals who had much higher oral competency in the paired task, therefore, were negatively impacted due to the performance of unsuccessful, communicative group work. In the end, the study questioned whether test takers with higher oral competency were treated unfairly in group work. The result is similar to the dilemma of the Cleary vs. Cole Model. Decisions based on paired-work or group performance may be unfair for some individual test takers.
Third, some empirical studies discuss the role of equal learning opportunities in ensuring test fairness (Walpole et al., 2005). It has been a concern that some test takers (e.g., minority) may have fewer educational opportunities and less access to quality resources, and thus are likely to perform poorly on the subsequent tests (Zwick, 2006). Although there is debate regarding whether opportunity to learn (OTL) is one aspect of test fairness as well as how to define OTL (Moss et al., 2005), empirical studies have been conducted examining the influence of learning opportunities on test scores. The literature on OTL generally found significant impacts of OTL on learning outcomes and test performance (Boscardin et al., 2002; Wang, 1998). Performance differences due to OTL were linked with various learner characteristics such as language, ethnicity, disability, and socio-economic status (Roach et al., 2009; Zwick, 2002). In the paper entitled *This test is unfair*, Walpole et al. (2005) investigated 227 urban African American and Latino high school students’ perceptions, test preparation, and information sources towards college admission tests such as Scholastic Aptitude Test (SAT). Findings showed that the African American and Latino students generally lacked information and resources to pay for tests and test preparation. As a result, they did not have adequate and equal opportunities to demonstrate their knowledge and perform at their best.

**Lack of Bias**

Bias refers to any source of construct irrelevant variance that may result in systematically higher or lower scores for groups of test takers (AERA, APA, & NCME, 1999). Systematic differences may be identified through statistical models (Camilli, 2006) or qualitative approaches (Fox, 2003; Karelitz, 2012), with the former used much more commonly. Many empirical studies have used psychometric approaches to
investigate statistical bias toward different groups (e.g., Chung & Berry, 2000; Lei, 2007; Young, 2001; Zwick & Sklar, 2005).

As stated earlier in the section on early development, attention shifted to investigate fairness for groups during the 20\textsuperscript{th} century. Subsequently, DIF research emerged in the 1980s, providing a psychometric method to determine if test items function differentially across groups of test takers (e.g., gender, ethnicity, and language) who were matched on ability (see Kunnan, 2000 and Ferne & Rupp, 2007 for comprehensive reviews of DIF research in language testing). Findings from DIF studies in language testing indicate that aspects of test items such as content and format as well as test takers’ background knowledge seem to be related to differential item performance between test taker groups (Bridgeman & Schmitt, 1997; Chung & Berry, 2000; Clapham, 1996). For example, Pae (2004) used the Mantel-Haenszel procedure and the IRT likelihood ratio approach to investigate English language test performance between test takers in the Humanities and Sciences. Pae found that seven items favoured test takers from the Humanities, and nine favoured the Sciences’ test takers. The preliminary content analysis of the test indicated that items dealing with science-related topics, data analysis, and number counting were differentially easier for the Sciences’ test takers, whereas items about human relationships were differentially easier for the Humanities. However, the existence of the flagged items was not sufficient evidence of bias.

In addition to using quantitative methods with the focus on score differences for bias detection, a few studies have examined perceptions of the fairness of the test as expressed by test takers and users, and/or employed qualitative approaches such as interviews and open response questionnaires to identify test bias (e.g., Fox & Cheng,
2007; Cheng & DeLuca, 2011). Du (2007) evaluated teachers’ perceptions of validity and fairness issues in a placement test through survey questionnaires and interviews. Results found that the test did not contain enough content knowledge to truly evaluate academic language proficiency, which led to unfair treatment regarding students’ placements. The study concluded that teachers’ perceptions helped to identify sources of construct under-representation. Fox and Cheng (2007) compared interpretations of two groups of test takers—22 students using English as a first language (L1) vs. 136 students using English as a second language (L2) in terms of test constructs and interactions among test design, interpretations, and accounts of classroom practices in a Canadian literacy test. Different perceptions regarding fairness were evident from the test takers’ comments, particularly in test accommodation, preparation, and administration for L2 students. In the study, focus group interviews proved to be a valuable tool to elicit stakeholders’ perceptions.

Karelitz’s study (2012) used surveys (n=11665) to examine public opinion on the validity and fairness of a standardized admission test. The majority of survey participants were test takers. Most participants perceived the test was unfairly biased against applicants from certain groups, despite the fact that statistical results had shown that the test did not discriminate against those groups including minorities, those with disabilities, or those from a lower socio-economic class. One valuable feature in terms of its research design is obvious — the researcher, also an official test evaluator, compared the results from two sides: information provided by stakeholders compared to psychometric evidence based on test scores. Through comparisons, the gap between the perceived fairness and evidence based on statistical results was found. The study concluded that the stakeholder involvement was useful since the participants (mostly test takers) provided
in-depth thoughts and meaningful insights regarding test fairness. The study suggested that a more dynamic channel of communication and dissemination of psychometric evidence between test developers and the public could be fruitful and beneficial.

Among the relevant research, Fox’s study (2003) is particularly insightful. Fox’s study used “an ecological approach for bias detection” to analyze test takers’ and raters’ responses with the intent to identify “bias that systematically undermines the measurement of the ability of a ‘group’ of test takers” (p. 26). The study compared accounts of test takers (n=423) and raters (n=12) on two versions (anchor vs. new) of a high-stakes language test through open response questionnaires. Results found that the new test version created fairness issues in some areas such as the test topics. Interestingly, those who expressed satisfaction or dissatisfaction with the topics were not of a particular gender, linguistic, cultural, academic background, or language proficiency group. The results indicate that group membership in bias research may not be easily identified. Unfairness may be related to any type of construct irrelevant variables from the perspective of test takers and raters. Ultimately, the perceptions and responses of test takers and raters on the new version resulted in the redefinition of the test specifications.

As Fox (2003) pointed out, the qualitative methods identified potential fairness concerns that might otherwise remain undetected by traditional methods such as DIF. Fox’s study (2003) has important implications in the fairness literature because it is one of the few studies that detected potential bias from the stakeholders’ perspectives (e.g., test takers) using a non-psychometric method, yet provided empirical information on test bias.
Equity

Equity, which is not synonymous with equality in this dissertation, requires that treatments are “appropriate and sufficient to the characteristics and needs of those treated” (Gordon & Bonilla-Bowman, 1999, p. xiv). In high-stakes testing, equity often implies that services and resources are to be distributed in proportion to need so test takers with special needs can perform at their best (Pitoniak & Royer, 2001). The dimension of equity is generally demonstrated in two areas: test accommodations and decision-making.

First, test accommodation practices have been widely studied (Finch, Barton, & Meyer, 2009; Moshinsky & Kazin, 2005). A primary purpose of test accommodations is to provide beneficial support to help disabled test takers to perform at their best since disability and learning difficulties interfere with their ability to demonstrate knowledge and skills. Test accommodations remove barriers, increase access, provide equitable opportunities, and offer a more accurate estimate of performance of these disabled test takers (Lang et al., 2005; Lazarus, et. al., 2009). Test accommodations generally do not change the construct being tested, and often involve special treatments such as extended time, large-type versions of tests, or read aloud of some or all of the items without changing the actual format of test items (Thurlow, Thompson, & Lazarus, 2006). Through meta-analysis, Li and Suen (2012) examined test accommodation practices for English Language Learners (ELLs). They found that test accommodations improved ELL’s test performance, but did not significantly influence the test performance of non-ELLs. These results indicated that test accommodations helped to remove barriers for
ELLs allowing them better demonstrate their knowledge and skills and produced no or negligible effects for non-ELLs.

Accommodation practices, which have been recommended over the last two decades, are increasingly perceived as an important and valuable aspect to promote fairness by treating test takers differently in test administration (Liu, 2010; Pitoniak & Royer, 2001). Some countries mandate the use of test accommodations for students with disabilities in high-stakes testing. In the United States, the Individuals with Disabilities Educational Act of 1997 was the first piece of legislation that required the inclusion of students with disabilities in assessment systems (Lazarus, et. al., 2009). However, test accommodations are not universally practiced and adopted, which may be related to different values in considering test fairness. Fan and Jing (2012) reported that there is generally a lack of provision of reasonable test accommodations in Chinese high-stakes testing programs. Empirical studies investigating test accommodation practices in China have not yet been found.

Second, in addition to test accommodations, admission decision-making in terms of equity practices has been researched. These empirical studies, mostly conducted in the United States, examined admission practices and value judgments embedded in admission practices (Alon & Tienda, 2007; Zwick, 2002; US Commission on Civil Rights, 2002). Breland et al. (1995, 2002) undertook longitudinal studies with institutes in the United States, and examined undergraduate admission policies, practices, and procedures through survey questionnaires. The study found that universities and colleges generally used a variety of measures for admission decisions. High school graduate grade point average and standardized test scores, such as the SAT, were the most widely used
factors in admission decisions. Results also found that some institutes adopted quotas to ensure representation and diversity of different racial groups in their student body, considering disparities in test scores between racial/ethnic groups especially in standardized tests. These institutes highlighted equity and granted extra “points” to minorities based on race in admission decisions. According to Gipps and Stobart (2009), the equity consideration in decision-making may also lead to testing policies and practices which allow different cut-scores for different groups in some testing contexts.

Ideally, if test takers all receive sufficient learning opportunities and they perform equally well on the criteria measure, they would have an equal chance of being chosen in decision-making, regardless of group membership (AERA, APA, & NCME, 1999). As a result, the fairness for each group and each individual may be achieved, similar to the Cleary Model. And admission decisions will be made according to the ranking. However, given a history of discrimination and inequality in learning opportunities as well as the under-representation of some ethnic groups in student bodies in the United States (e.g., Black and Hispanic test takers), race-sensitive admission criteria have been practiced in some universities and colleges (Breland, 2002; US Commission on Civil Rights, 2002). For these institutes, fair selection implies equitable treatment in decision-making, which requires an adjustment in accept-reject rates for different groups so that every group has certain proportion of successful applicants in student bodies.

Nevertheless, in the studies of Breland et al. (1995, 2002), not all universities and colleges used race as a criterion for decision-making. In these universities, admission decisions were primarily based on ranking, regardless of group membership. The 2002 study is particularly interesting because it was conducted after racial quotas were
prohibited in some states, such as Taxes and California. In these states, it was ruled as unconstitutional for public colleges and universities to use race as a condition of admission. The abandonment of racial quotas significantly impacted enrollment of under-represented groups (e.g., African American, Hispanics, and Native Americans) in these states, which was questioned by some researchers from a fairness perspective (Zwick, 2002).

The dilemma of balancing equity and equality in decision-making has been debated for decades. As described earlier, measurement specialists did not agree on a preferred theoretical model (Cleary, 1968; Cole, 1973). They believed that fair selection is linked to subjective value judgments (Petersen & Novick, 1976), which remains true in the measurement area (McNamara & Roever, 2006). In practice, whether equitable, race-sensitive considerations should be used to decide admission decisions through adjustment in accept-reject rates for different groups has received academic, judicial, and legislative challenges (US Commission on Civil Rights, 2002). A recent legal dispute in April 2013, Fisher v. University of Texas, took place when Abigail Fisher, a white student, questioned whether or not race should play a factor in admission policies at University of Texas. In January 2014, the California Senate discussed the potential adoption of Senate Constitution Amendment No. 5 (SCA 5), which would repeal the previous provisions and allow California to deny an individual or group's rights to public education on the basis of race, sex, color, ethnicity, or national origin. Under this legislation, public colleges and universities in California could use race as a condition of student admission. As discussed in the special issue of Educational Researcher (April, 2013), such legal and legislative disputes reflect conflicts of equality, equity, and diversity in fair decision-making in
higher education. Whether equitable and race-sensitive decision-making should be adopted and practiced to ensure fairness in admission decisions remains an unfinished research agenda.

*Comparability across Different Forms*

Comparability can be defined in different ways. In this section, I use the term comparability to refer to the examination of the commonality of score meaning across different forms as well as testing conditions such as delivery modes and computer platforms (Bridgeman & Schmitt, 1997). Comparability across parallel and modified forms in the field of measurement is to ensure that test takers are provided with equal opportunities to demonstrate acquired knowledge and skills, regardless of test forms or conditions. Various classical and IRT equating procedures have been tested and used to make necessary adjustments so test takers are neither advantaged nor disadvantaged by the particular test form they take (Cook & Eignor, 1991; Dorans, Pommerich, & Holland, 2007). These equating techniques ensure that the test scores from different forms are equivalent, help to compare test taker performance, and track performance growth across time intervals (von Davier, 2007). Researchers have also investigated modified testing forms for certain groups of test takers such as those with special needs (Abedi, 2009; Cawthon, Highley, & Leppo, 2011; Elliott & Roach, 2007). In addition, potential construct irrelevant factors such as comparability between paper-based and computer-based tests, computer familiarity and proficiency, platform familiarity, and user interface have been examined (Bridgeman, 1992; Choi, Boo, & Kim, 2003).

In summary, empirical studies have been conducted to investigate various dimensions of test fairness, and the specific results of these empirical studies are varied.
The review of the empirical studies reflects the multiple dimensions of test fairness: equality, lack of bias, equity, and comparability. Fairness investigations have expanded from an initial focus on equality for all test takers to issues of bias, equity, and comparability toward test taker groups. Overall, the current empirical studies generally recommend that test fairness requires that individuals and groups of test takers take the same or equivalent tests, under the same or accommodated conditions, and their performances are evaluated using the same scoring schemes. However, it is unclear whether these generally agreed practices derived from the empirical studies, mostly conducted in the West, are accepted universally in different contexts (e.g., equity in test accommodations). In addition, the fairness in score-based decision-making remains controversial (Breland et al., 1995, 2002). Proposals that seem to improve a test’s fairness from one perspective may have counterarguments from another.

The review of empirical studies also provides important implications regarding research methods and participants. While statistical models and quantitative techniques (e.g., DIF approaches, equating methods, and selection models) are used to detect bias and ensure comparability, fairness studies have also used qualitative methods such as interviews, with a wide range of test stakeholders including test takers, teachers, proctors, program administrators, raters, and educators. Overall, these empirical studies typically investigated one specific aspect of fairness within a testing context. It is inevitable, when examining one specific aspect of fairness, that the studies tend to offer limited insight into the scope of test fairness. What is needed is a more comprehensive examination regarding the scope of fairness investigations. In the following, I review two classes of conceptual work—standards, guidelines, and codes of practices as well as individual
scholarly work, which will provide a thorough description of the scope of test fairness to
guide my dissertation.

Testing Standards, Guidelines, and Codes of Practices

Testing standards, guidelines, and codes of practices are developed by large committees or testing publishers to provide guidance on fairness practices for the broader educational communities. Generally, fairness has been given a prominent position in these materials (JCTP, 2004; ETS, 2002). Standards, Guidelines, and Codes of Practices provide valuable information to understand the concept of test fairness and its role. My selection of testing standards, guidelines, and codes of practices was based on the criteria of usefulness and relevance. Due to my topic on the fairness of the GSEEE — a large-scale high-stakes, external language test, I chose documents that were generally developed by professional organizations, commercial test publishers, ministries and departments of education, and other educational systems. Since fairness in high-stakes testing may be different from fairness in classroom settings (Gipps & Stobart, 2009), the standards, guidelines, and codes of practices that examine only classroom teacher-designed testing, for example, the Student Evaluation Standards: How to improve evaluation of students (the Joint Committee on Standards for Educational Evaluation, 2002), were not included. I also focused on documents in educational testing and assessment with special attention on language testing and assessment due to my research context.

A multi-database search strategy was used to identify documents. Reference citations were identified via: 1) screening of reference lists in scholarly educational and language testing and assessment journals written in English between 1990 and 2012; and
2) international academic databases such as Educational Research Complete, ERIC, and Proquest Digital Dissertations as well as technical reports published by testing companies such as Educational Testing Service (ETS) in the United States. My search identified 15 documents written by 11 group authors (see Table 1). Some of these documents only have one version, while others, such as the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999; hereafter called the 1999 Standards), have been revised over time. In this case and cases like this, the most recent revision was used.

Table 1

*A list of Testing and Assessment Standards, Guidelines, and Codes of Practices*

<table>
<thead>
<tr>
<th>Group authors</th>
<th>Year of Publication</th>
<th>Standards/Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Committee on Testing Practices (JCTP)*</td>
<td>2004</td>
<td><em>Code of Fair testing practices in Education Rights and responsibilities of test takers: Guidelines and expectations</em></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Joint Advisory Committee (JAC)*</td>
<td>1993</td>
<td><em>Principles for Fair Student Assessment Practices for Education in Canada</em></td>
</tr>
<tr>
<td>Australasian Curriculum, Assessment and Certification Authorities (ACACA)</td>
<td>1995</td>
<td><em>Guidelines for Assessment Quality and Equity</em></td>
</tr>
<tr>
<td>Educational Testing Service (ETS)</td>
<td>2002</td>
<td><em>ETS Standards for Quality and Fairness</em></td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td><em>ETS Guidelines for Fairness Review of Assessment</em></td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td><em>ETS International Principles for Fairness Review of Assessments</em></td>
</tr>
<tr>
<td>International Language Testing Association (ILTA)</td>
<td>2007</td>
<td><em>Guidelines for Practice</em></td>
</tr>
<tr>
<td>Association of Language Testers in Europe</td>
<td>1994</td>
<td><em>Code of Practices</em></td>
</tr>
</tbody>
</table>

---

3 The JCTP represents 7 member associations in the United States.
4 The JAC represents the 8 professional associations in Canada.
While each of these documents specifically addresses issues related to fairness, the relative importance of fairness varies across the documents. In some documents (e.g., JCTP, 2004), fairness is conceptualized as the overriding concern for all aspects of assessment. In contrast, other documents (e.g., ALTE, 1994) do not consider test fairness as the central issue. Essentially, I used these documents to pinpoint key fairness issues, which helped me to identify a conceptual framework that would be appropriate to guide my dissertation in a Chinese testing context. I identified not only the commonalities across these documents but also the less common features within the documents. The inclusion of less common features is important because it enabled me to discover issues that were not congruent but have led to much of the ongoing discussion and debate in the study of test fairness. Overall, the identified standards, guidelines, and codes of practices primarily reflect three major issues of test fairness:

1) Which stakeholder groups are included/excluded in the discussion of test fairness?

2) What is the scope of test fairness under discussion?

3) How is fairness related to validity?

---

Stakeholder Groups

Generally, these documents identify three groups of stakeholders—test developers, test users, and test takers. However, the documents define these groups, especially the groups of test developers and test users in different ways. For example, the *Code of Practice* (ALTE, 1994) defines test developers as “people who actually construct and administer examinations as well as those who set policies for particular testing programmes,” and test users as “people who select examinations, commission examination development services or make decisions which affect the educational possibilities and careers of others based on the basis of examination results” (p. 1). In comparison, the *1999 Standards* (AERA, APRA, & NCME, 1999) considers the test developer as “the person(s) or agency responsible for the construction of a test and for the documentation regarding its technical quality for an intended purpose” and the test user as “the person(s) or agency responsible for the choice and administration of a test, for the interpretation of test scores produced in a given context, and for any decisions or actions that are based, in part, on test scores” (p.183). Clearly, test administrators are classified as test developers in the *Code of Practice* while defined as test users in the *1999 Standards*. The difference may be a matter of context. In some testing situations, test developers oversee administration, while in others test users manage test administration. Although not a substantive difference, there may exist contextual differences in these conceptualizations. Such differences indicate that it is important to clarify groups of test stakeholders in a specific testing context since each stakeholder group has different rights and responsibilities.
Furthermore, while some documents focus on obligations and professional conduct of test developers, a majority of the documents also include rights and responsibilities of test users and/or test takers to ensure fairness. For example, the *Guidelines for Good Practices in Language Testing and Assessment* (EALTA, 2006) and *Guideline for Assessment Quality and Equity* (ACACA, 1995) are largely intended for test developers. In contrast, the *Code of Good Testing Practices* (JLTA, 2002) and the *Code of Practice* (ALTE, 1994) are intended both for test developers and users. The 1999 *Standards* (AERA, APA, & NCME, 1999) and the *Guidelines for Practice* (ILTA, 2007) appear to include all three groups: test developers, users, and takers. Some documents focus on the role of the ethical, professional conduct of test developers to ensure fairness (ACACA, 1995; EALTA, 2006). In such documents, what constitutes fair practices is largely left to testing experts who have professional training and expertise in test design and development as well as item analysis and test evaluation. In comparison, other documents add the roles of test users and test takers and provide clarifications about rights and responsibilities of these groups (AERA, APA, & NCME, 1999; JAC, 1993). Fairness in this approach is portrayed to be in the hands of all those who are involved in the testing situation. Such an approach provides the potential to encourage and promote inclusivity and interaction among different groups of stakeholders. This is an important notion as the voices of test users and test takers tend to be less valued in the conversations of what constitutes fairness (Rea-Dickins, 1997).

*The Scope of Test Fairness*

One of the important contributions of standards, guidelines, and codes of practices is that they generally provide a clear scope of issues for fairness investigations within the
document, although the scope may be inconsistent from one document to another. A majority of the standards and codes of practice limits the scope to testing activities or roles of the involved parties across the testing process (e.g., EALTA, 2006; JAC, 1993; JCTP, 2004). Within these bounds across the testing process, my amalgamation of the documents identified four major areas: test design and development, administration, scoring, and score-based test use. Apart from these common areas, two other less common areas were identified: opportunity to learn (OTL) and impact of test use. Only two documents emphasized opportunity to learn (AERA, APA, & NMCE, 1999; ACACA, 1995) and four discussed the importance of positive impact of test use (ALTE, 2001; EALTA, 2006; ETS, 2005; University of Cambridge, 2011). Combined, these areas appear to provide a comprehensive picture regarding various views on the scope of test fairness. In the following, I discuss the testing process that includes four common areas, followed by two less common areas.

The Testing Process

The standards, guidelines, and codes of practice illustrate procedures to address issues of fairness across the testing process. To clarify, the testing process includes activities directly related to testing including (1) test design and development, (2) administration, (3) scoring, and (4) score-based test use. It excludes what precedes a test as well as its social consequences.

Test design and development. In the first area of test design and development, the standards, guidelines, and codes of practice are generally consistent with respect to the importance of practices such as selecting and training item writers, pretesting/field testing, evaluating item performance (i.e., Differential Item Functioning), conducting
fairness/sensitivity reviews, and examining comparability across parallel and modified forms. In general, fairness in test design and development puts much emphasis on the examination of group differences. There is consensus among the documents that a lack of bias and comparability across different forms are key elements for providing fairness in test design and development. Test specifications describing test characteristics such as test content, format, and length should take all test takers into consideration and avoid disadvantaging certain groups of test takers. Choices of what to test among competing constructs and the formats and forms to test the content are important aspects to decide if groups of test takers will have fair opportunities to perform. By measuring narrowly defined construct(s) or introducing irrelevant variance, a test may underestimate (or overestimate) the competence of groups who would have done better (or poorer) on what was not included (Willingham & Cole, 1999).

Administration. In the second area—test administration, common fairness practices address the need for consistent administration procedures, test accommodations, and security. These procedures must also be designed to minimize cheating. There are two foci in test administration. On one hand, all the documents state that test administration, such as standard time limits, proctoring to ensure there are no irregularities, and testing physical conditions, should be as identical as possible for all test takers, taking any form of the test at any time and location. Standardized administration is desired to ensure a test is administered under uniform conditions unless there are defensible reasons to adjust the administration procedures. On the other hand, the majority of the documents point out the importance of special accommodations. Test accommodations, which imply different treatments during test administration, are
considered to be crucial to ensure test fairness (e.g., ILTA, 2007; JCTP, 2004; University of Cambridge, 2011). However, not all the documents openly promote and encourage the practice of test accommodation. For example, the *Code of Good Testing Practices* (JLTA, 2003) and the *Guidelines for Good practice* (EALTA, 2006) completely avoid accommodations and do not discuss any accommodation practices. This result is consistent with the review of empirical studies – test accommodations are not universally practiced and adopted (Fan & Jing, 2012). There is a potential that different testing contexts may define differently regarding whether fair testing practices should include test accommodations.

**Scoring.** The third area in which fairness issues must be considered is scoring. The standards, guidelines, and codes of practice generally agree that fairness requires impartiality and consistency in scoring for all test takers. Fair scoring implies that all test takers be treated equally with the same scoring procedures and schemes. Common fairness practices in scoring include standard scoring procedures, raters’ training, score communication/reporting, and appeal systems if there are any concerns about test scores. The scoring procedure and rater’s training are important to maintain accuracy and consistency of test scores.

**Score-based test use.** Lastly, the documents broadly point out the importance of fairness in the use of scores. Fairness in test score use includes considerations such as recommended interpretations, appropriate uses and communication, deterring misuse, encouraging multiple sources in decision-making, standard errors of measurement of the test, and monitoring the extent to which the test is fulfilling its intended purposes. Only two documents provide specific information as to what constitutes fairness in decision-
making: the 1999 Standards (AERA, APA, & NCME, 1999) and the Principles for Fair Student Assessment Practices for Education in Canada (JAC, 1993). While these two documents explicitly explain the meaning of fairness in test score use, they take opposite stances. The 1999 Standards (AERA, APA, & NCME, 1999) stresses equality in the outcomes of testing, indicating that test takers who perform equally well on the criterion measure should have an equal chance of being selected, irrespective of group membership. In contrast, the Principles for Fair Student Assessment Practices for Education in Canada (JAC, 1993) specifies that test users should “examine the need for local passing or cut-scores and, if called for, reset these scores” (p.18). The underlying notion is that cut-scores and decision-making may be different for students with the same score due to their different group membership (e.g., geographical locations). The contradiction between the 1999 Standards and the Principles for Fair Student Assessment Practices for Education in Canada reflects different conceptualizations of the fairness in decision-making. Fundamentally, such discrepancy indicates the ongoing debates regarding fairness in score-based test use. It is consistent with the measurement, empirical, and legal debates on fair decisions as demonstrated earlier (e.g., the Cleary vs. Cole model). Once again, what is considered as fair decision-making is related to value judgments.

Opportunity to learn (OTL)

Among the 15 documents, two documents discuss the area of OTL in order to achieve test fairness. The 1999 Standards (AERA, APA, & NCME, 1999) points out four prevalent features of test fairness and one of them is sufficient opportunity to learn prior to achievement testing. The 1999 Standards (AERA, APA, & NCME, 1999) specifies the
importance of providing opportunities to learn the material covered in achievement tests since those tests are intended to assess what test takers know or can do as a result of formal classroom learning. The second document, *Guidelines for quality and equity* (ACACA, 1995) also explicitly discusses OTL, but takes a different stance:

ACACA agencies are responsible for assessment, not for how schools operate. The need to be fair to all students means that ACACA assessment practices should not seek to adjust for missed opportunities to learn by saying that students have demonstrated achievement that they have not in fact demonstrated. In this sense the equity issues connected with ideas of opportunity to learn are not, therefore, part of these guidelines. (p. 1)

OTL, at least in achievement testing, is considered to fall within the scope of fairness discussions in the *1999 Standards*. In contrast, the Australian document (ACACA, 1995) considers OTL to be beyond the scope of fairness investigations. Such divergence is consistent with the review of empirical studies. Overall, OTL is not commonly discussed in the standards, guidelines, and codes of practices. In the two documents in which it is discussed, the views are quite divergent.

*Impact of Test Use*

The impact of test use is frequently discussed in test validation research (Cheng, 2008). Four documents, all in language testing, argue that a fair test has to bring positive impacts to teaching and learning and society in general to accomplish fairness of a test (ALTE, 2001; EALTA, 2006; ETS, 2005; University of Cambridge, 2011). For example, the *Principles for Good Practices by ALTE Examinations* (ALTE, 2001) defines fairness as an overriding concern in assessment from test design to test impact on educational
processes and society in general. This document states that “examination developers must cooperate with the aim that their examinations will not have a negative impact and, as far as possible strive to achieve positive impact” (p. 13). A fair test, by this view, should not be harmful or detrimental to learning, teaching, and society.

While the four documents in language testing argue that fair test use has to also evaluate the positive and negative impact of testing, others restrict the meaning of fair test use to the inferences that have been, and can be expected to made from test scores and the implications of the effects of decisions made using these scores (e.g., ILTA, 2007; JCTP, 2004). Again, whether test fairness should consider impact of test is related to the different views regarding the scope of fairness investigations. Since consequences of test use are influenced and mediated by many factors including the test, test takers, teachers, and educational settings, significant challenges remain about which factors should be investigated and under which conditions beneficial impacts are most likely to be generated (Cheng, 2008).

*Fairness and Validity*

Fairness has been conceptualized in various ways and one of the central issues is how fairness is related to validity (Xi, 2010). Three different approaches have been used to depict the relationship between fairness and validity across the investigated documents. The first approach views test fairness as an overarching test quality that consists of different aspects including validity. The *Principles for Fair Student Assessment Practices for Education in Canada* (JAC, 1993) and the *Code of Fair Testing Practices in Education* (JCTP, 2004) state that fairness is a primary consideration in all aspects of testing and evidence in terms of technical qualities such as reliability and validity are part
of fair testing practices. Similarly, the *Principles of Good Practice for ALTE Examinations* (ALTE, 2001) starts with the statement that fairness is an overriding concern in all aspects of assessment. At the beginning of the *Guidelines for Practice* 6 (ILTA, 2007) and *Code of Good Testing Practices* (JLTA, 2002), it is highlighted that validity and reliability are two basic considerations for good testing practice in all situations. The two documents then illustrate how fairness issues are integrated into each stage of the testing process. This approach can be found in the work of some researchers such as Kunnan (2008). Hence this conceptualization of fairness implies that test fairness is overarching and subsumes reliability and validity.

The second approach perceives fairness as a separate concept from validity. The two terms, validity and fairness, exist side by side; however, how validity and fairness connect with each other remains unclear. The *Principles of Good Practice* states the principle of high-quality assessments is to ensure that “assessments are valid and that all test takers are treated fairly” (University of Cambridge, ESOL Examinations, p. 8). In The *ETS Standards for Fairness and Quality* (ETS, 2002), validity and fairness are put into two chapters. Although some tenuous references may be made to validity, ETS (2002) emphasizes equality and the importance of taking groups of test takers into consideration. It considers fairness as “the extent to which the test is appropriate for members of different groups, and the extent to which test takers are treated the same way, regardless of gender, race, ethnicity, and the like” (p. 71). This second approach seems to separate validity and fairness, acknowledging that fairness and validity are both

---

important. Nonetheless, this approach does not clarify the actual differences between the two concepts.

The third approach considers fairness and validity to be closely interrelated (ACACA, 1995; AERA, APA, & NCME, 1999). In the 1999 Standards, validity evidence such as construct underrepresentation, construct irrelevant variables, test takers’ or raters’ response processes, relations of test scores to other criteria, and uses of testing is discussed first. Then the corresponding fairness issues for each type of validity evidence are clarified and expanded in a separate chapter, with some attention given to different subgroups of test takers. This connection between discussions of fairness and validity suggests “a strong possibility for linking fairness with validity in a principled way” (Xi, 2010, p. 152). Major issues discussed in validity are also reflected in various aspects of fairness investigations. The approach is agreeable to scholars such as Willingham and Cole (1997), Gipps and Stobart (2009), Xi (2010), and Kane (2010).

While acknowledging that validity traditionally focuses more on within-group variation and fairness on between-group variation, this approach considers issues being addressed under validity and fairness to be primarily the same (Camilli, 2007; Kane, 2010).

As stated in Chapter 1, I believe fairness is highly intertwined with validity and this view is closest to the third approach. The purpose of both validity and fairness evidence is to ensure that a test is designed accurately and used appropriately for its intended purposes. The interrelatedness of fairness and validity is particularly evident in examining fairness for individuals because individual fairness is “proportional to, if not synonymous with, validity” (Willingham, 1999, p. 220). Both fairness and validity examine the extent of errors of estimates or inference in test scores. The smaller the
errors of estimate or inference, the more valid the interpretations of tests and the fairer the tests are. As Willingham (1999) contended, “Improved validity is the road to fairness” (p. 221). As such, the two terms are closely related. However, as Kane (2010) states, fairness and validity may discuss “somewhat different points of view and involve different emphasis” (p. 78). While validity examines score-based inferences and decision-making (Kane, 2001), fairness focuses on comparable treatment for test takers, both groups and individuals (Willingham, 1999).

In conclusion, the review of the standards, guidelines, and codes of practices offers important, complementary information regarding test fairness in three aspects: the involvement of various stakeholder groups, the scope of test fairness, and fairness and validity. As reviewed, the inclusion of various stakeholder groups has implications to better understand the complexity of test fairness. The synthesis of the standards, guidelines, and codes of practice also provides unique contributions regarding the relationship between fairness and validity. Finally, the standards, guidelines, and codes of practices offer important, yet inconsistent information regarding the scope of test fairness. The scope of fairness investigations needs to be clarified in fairness research. In the following, I review the more recent, insightful, individual researchers’ conceptual work on test fairness with the intent to identify an appropriate scope and a conceptual framework to guide my research.

Individual Conceptual Work

This section reviews the major individual researchers’ conceptual work related to test fairness, with a special focus on those pieces that were written in the new millennium. These conceptual works provide more updated, original information with
much elaboration. My search and reading led me to six conceptual frameworks that inform high-stakes test fairness in educational and language testing and assessment. These include the works of Willingham (Willingham & Cole, 1997; Willingham, 1999), Xi (2010), McNamara (McNamara & Roever, 2006; McNamara & Ryan, 2011), Camilli (2006, 2013), Gipps and Stobart (Gipps & Stobart, 2009; Stobart, 2005), and Kunnan (2004, 2008). Broadly speaking, these works can be categorized into two types: one focusing on the testing process and the other highlighting the socio-cultural contexts. Although these conceptual works all recognize the role of both the testing process and socio-cultural contexts in investigating test fairness, their emphases are different. While some of these conceptual frameworks primarily focus on the testing process from test design to score-based use, some also consider what precedes testing and the subsequent consequences of testing and highlight the influence of the socio-cultural contexts. In the following section, I explain how the focus of the two perspectives differs. By elaborating on these two different perspectives, my intention is to facilitate the identification of an appropriate scope to guide my dissertation research. The one that is more applicable was selected to guide this dissertation research.

*The Perspective Focusing on the Testing Process*

This perspective focuses on fair treatment of test takers in various aspects of the testing process, from test design, development, administration, scoring, to score-based use. In other words, a test is fair when test takers are provided with appropriate, sufficient opportunities to perform at their best within the testing process. The scope of this perspective addresses any potential advantages or disadvantages in test items and alternative tests toward certain test taker groups (e.g., DIF, parallel and modified forms),
consistent and secure test administration procedures (e.g., standardized administration), proper accommodations for disabled test takers, impartial scoring, and appropriate decision-making. This perspective generally focuses on “the technical quality of the test” (McNamara & Ryan, 2011, p. 161). It limits the scope of fairness investigations within the aspects that are linked to actual, direct testing activities. This perspective is endorsed by Willingham and Cole (1997), Xi (2010), and McNamara and Ryan (2011).

Among all of these conceptual works, Willingham and Cole’s work (1997) is the first book devoted exclusively to test fairness. It provides a well-argued framework regarding test fairness. Their fairness framework included “comparability of opportunity for examinees to demonstrate relevant proficiency, comparable assessment exercise and scores, and comparable treatment of examinees in test interpretation and use” (p. 11). Willingham and Cole stated that test fairness is best conceived as comparability for all individuals and groups from test design, development, administration, and use. They argued that comparable opportunity to demonstrate skills and knowledge is not the same as comparable opportunity to acquire skills, and test fairness only addresses the former. Willingham’s later work (Willingham, 1999), a chapter in the book Assessment in Higher Education that was dedicated to Willingham and edited by Samuel Messick (Messick, 1999) provided a systematic view of test fairness. In this work, Willingham (1999) expanded this framework into two levels: the test fairness manifold that included the participants, process, and situation, as well as the social matrix that covers fairness, usefulness, and practicality. The fundamental issue in this extended framework, however, remains the same: fairness is about comparability investigations for groups and individuals of test takers from test design to score-based test use. Willingham stressed
that “group differences derived from wide disparities in education and social conditions—as well as personal experience and interests—do not necessarily reflect test unfairness” (1999, p. 220). Although Willingham believed that the measurement profession should be actively engaged in the discussion of social justice and socio-cultural influence, he limited the scope of test fairness to direct testing activities. Overall, Willingham’s work contended that a fair test should be comparable from person to person and group to group across all aspects of the assessment process from the initial design to score-based test use.

Xi (2010) echoed Willingham’s work, highlighting the importance of the testing process in fairness investigations from assessment conceptualization to the use of assessment results. Xi’s work adopts the perspective focusing on fair treatment across the testing process from test design to use, disregarding learning opportunities. Xi’s work differs from Willingham’s conceptualization in that she argued fairness only for identifiable and relevant groups. To “plan and set priority in fairness investigations” (p. 160), Xi contended systematic, demonstrable evidence at the group level, disregarding the study of individual differences in fairness research. Based on Kane’s validity argument (1992), Xi developed the “fairness argument” to investigate test fairness for test taker groups. As pointed out earlier, test fairness is derived from the pursuit of equality among all individuals. Although fairness investigations have been shifted on group differences since the 20th century, fairness for individuals continues to be promoted and examined in the literature (ALTE, 2001; May, 2009). The challenge remains in terms of how tests are designed, administered, and used to address group differences without losing the benefits of standardized testing based on equal treatment for all individuals.
In summary, the perspective focusing on the testing process highlights the testing process and technical properties of the test (McNamara & Ryan, 2011; Willingham, 1999). This perspective generally examines actual testing activities related to direct testing actions in test design, development, administration, scoring, and score-based use. It examines the testing process and largely situates itself in psychometric practices and cognitive science. This perspective separates test takers from their educational background, experience, and social consequences in discussing test fairness. In this view, test takers’ learning opportunities and social consequences of test use are beyond the scope of fairness investigations, for example, possibly falling into the scope of social justice (McNamara & Ryan, 2011).

The Perspective Focusing on the Socio-cultural Context

In contrast, the second perspective highlights the importance of socio-cultural implications with a broader scope of test fairness investigations from what precedes testing, the testing process, to the subsequent consequences of testing. According to this view, test fairness may be defined as justice and beneficence (Kunnan, 2008) and equity (Gipps & Stobart, 2009). This perspective contends that knowledge and skills are not only related to mental function and information processing within test takers’ minds, but also inseparable from interactions between test takers and their larger learning environments as well as social goals such as diversity in student population. This perspective stresses an extended scope of fairness investigations, in particular, opportunity to learn. The focus of this perspective shifts toward analyses of the socio-cultural context, which may include many aspects and various social, cultural,

---

7 Compared to the definition of equity in this dissertation, the word is used differently in Gipps and Stobart’s work. They used equity broadly, referring to “moral justice” and “the spirit of justice” (Secada, 1989, p.81).
educational, political, legal, economic, and philosophical considerations. Investigations using this perspective rely on analyses such as context analysis, cultural analysis, and critical policy analysis (Moss et. al., 2005).

Gipps and Stobart argued that fair assessment should take the context of assessment and social and cultural issues into consideration, in addition to technical aspects (Gipps & Stobart, 2009; Stobart, 2005). They believed that fairness in assessment involves “both what precedes an assessment (for example, access and resources) and its consequences (for example, interpretations of results and impact) as well as aspects of the assessment design itself” (Gipps & Stobart, 2009, p. 105). They used equity interchangeably with fairness, which they believed is socially embedded and can only be fully understood by taking account of the social and cultural contexts along with the technical characteristics. Gipps and Stobart used examples to illustrate the complexity of fairness in different social, cultural, and assessment contexts. For example, prestigious universities in the United Kingdom adopted the preferential admission practice to students studying in public high schools over those from private high schools. Although they did not further elaborate on how specific social, cultural, political considerations influenced testing practices and admission decisions, their work stresses the important role of the socio-cultural context and the larger scope of test fairness. They believed that fairness and equity cannot be assumed, but must be carefully investigated in any assessment environment by looking at its social and cultural context.

Similar to Gipps and Stobart’s work, Kunnan’s work emphasized the importance of value implications and social consequences (2000; 2004; 2008). His most recent paper
(2008) proposed a test fairness framework (TFF) and test context framework (TCF) based on the two principles of justice and beneficence:

The Principle of Justice: A test ought to be fair to all test takers, that is, there is a presumption of treating every person with equal respect.

The Principle of Beneficence: A test ought to bring about good in society, that is, it should not be harmful or detrimental to society. (p. 231)

The TFF includes five main qualities in the testing process: validity, absence of bias, access, administration, and social consequences. In addition to the TFF, Kunnan (2008) highlighted “the socio-economic-political issues that are critically part of tests and testing practices” (2008, p. 14). He elaborated on the TCF that includes four aspects: 1) the political and economic, 2) the educational, social and cultural, 3) the technological and infrastructure, and 4) the legal and ethical considerations. Kunnan’s work implies a wide range of discussions and potential opportunities to explore various contextualized considerations regarding test fairness from a multiple-disciplinary perspective.

The perspective focusing on the socio-cultural context is also echoed by Camilli (2006, 2013). Camilli contended “the evaluation of fairness requires a broad range of evidence that include empirical data, but may also involve legal, ethical, political, philosophical, and economic reasoning” (2006, p. 225). He described how the social and legal contexts of the United States, such as the Fourteenth Amendment, shape fairness issues and judgments. The Amendment, as he explained, promotes equality, which indicates treatment should be as identical as possible and prohibits discrimination and disparate impact to individuals. However, the Amendment also states the importance of the protection of certain designated groups (e.g., race, ethnicity, and disability) in the
United States since these groups may not enjoy the same learning opportunities and cultural advantages. Camilli pointed out this brings difficulty to individual fairness—“discrimination might be demonstrable statistically at the group level, but not in the case of any particular individuals” (2006, p. 223). In his recent work, Camili (2013) further discussed that a classification of designated groups is often socially constructed, and fairness is framed by moral and ethical choices in a specific context.

Overall, central to this perspective is the philosophical and epistemological underpinning that highlights legal, social, cultural, and political factors influencing test fairness. The fundamental issue is, according to this perspective, that a test is fair if it is mediated successfully through historical, social, cultural, moral, and political lenses with acceptance and endorsement among test stakeholders and in society. Fair testing does not just include practices to provide test takers with comparable opportunities to demonstrate their knowledge and skills within the testing process, but also needs to be culturally, socially, politically, and ideologically acceptable. The attention shifts from investigating the testing process to context analysis with the focus on how a particular context shapes the complexity and dynamics of test fairness. Context analysis would ask for a wider disciplinary net within anthropology, sociology, political science, psychology, linguistics, and many more (Haertel, et. al., 2008; Moss et. al., 2005). The intent is to identify the contextual forces (e.g., technical, legal, socio-cultural, political, educational, ideological, and economic forces), which are critical in framing test fairness. Such multidisciplinary work needs the study of the context, in particular, related to various forces that have shaped the current practices or seek to change the current practices (e.g., Lewin’s force-field analysis, 1943). This perspective requires genuine interactions among all those who
are involved in testing to have in-depth investigations of alternative, possibly competing, views on test fairness.

Summary of the Literature

In conclusion, this literature review synthesizes various conceptualizations of test fairness from four sources. Each source has identified relevant issues regarding the complexity of test fairness. While the early development and empirical studies provide insight into multi-dimensionalities of test fairness (e.g., equality, lack of bias, and equity), the standards, guidelines, and codes of practice offer directions about the scope of fairness investigations. The standards, guidelines, and codes of practice also help to understand the importance of stakeholder involvement as well as the relationship between fairness and validity. Finally, the individual conceptual work provides the most recent, critical views on test fairness. Two broad conceptual perspectives are identified, one focusing on the testing process and the other on the socio-cultural context. The following are the seven major conclusions derived from the four literature sources:

1. The literature agrees that test fairness is larger than, and inclusive of, the early view of equality—all test takers should sit the same test under the same conditions with the same scoring procedure.

2. The literature agrees that test fairness is larger than, and inclusive of bias (AERA, APA, & NCME, 1999; McNamara & Rover, 2006).

3. The recent fairness literature generally agree that test fairness requires that groups and individuals of test takers take the same or equivalent tests, under the same or accommodating conditions, and their performances are evaluated under the same scoring procedures (NCME, 1995; Willingham & Cole, 1997).
4. The recent fairness literature generally agrees that fairness in decision-making remains controversial and is often related to value judgments (JAC, 1993; Camilli, 2006).

5. The recent fairness literature agrees that test fairness is multi-dimensional and involves multiple stakeholder groups across multiple aspects of the testing process from test design to score-based use (Breland et al., 2002; AERA, APA, & NCME, 1999).

6. More recently, the fairness literature highlights the importance of analyzing the socio-cultural context, and argues for a broad scope of test fairness including not only the testing process but also opportunity to learn and social consequences of testing (Gipps & Stabort; 2009; Kunna, 2008; Moss et. al., 2005).

7. Most fairness literature views that fairness and validity are related, although conflicting views exist as for which has a larger scope.

Overall, the review of the four sources sheds important insights into the method for action taking (e.g., the stakeholder involvement) and possible directions (e.g., the focus of the testing process). However, ambiguity, controversies, and complications remain, especially when researchers use already convoluted concepts such as social justice to explain another complex concept: test fairness. Although the recent conceptualizations generally agree that test fairness requires that groups and individuals of test takers take the same or equivalent tests, under the same or accommodating conditions, and their performances are evaluated based on the same criteria, it remains unclear whether they are endorsed in all testing settings, for example, the GSEEE.

Further, what constitutes fair treatment in decision-making involves subjective value
judgments. These value judgments generally remain under-researched in Chinese testing programs. There is a need to examine how fairness is defined in large-scale high-stake testing such as the GSEEE.

Willingham’s Conceptual Framework and Operational Model

There are two parts of this section. In the first part, I provide a rationale as for why Willingham’s framework was chosen as an overarching framework in this dissertation, and elaborate on four key elements of his conceptual framework (Willingham & Cole, 1997; Willingham, 1999). Willingham is a major contributor to the fairness framework (Willingham & Cole, 1997) and the Test Fairness Manifold (Willingham, 1999). Both works were used to guide this dissertation. In the second part, I establish a specific model to guide my dissertation study. The model provides a visual representation of the conceptual framework and illustrates the detailed components to guide fairness investigations in action.

Willingham’s Conceptual Framework

The conceptual framework that guides my doctoral research is based on the work of Willingham, both his fairness framework (Willingham & Cole, 1997) and his Test Fairness Manifold (Willingham, 1999). There are two reasons that I chose Willingham’s work as an overarching framework. First, Willingham’s work provides “a solid description of the complexity of the issues surrounding fairness and assessment” (Welch, 1998, p. 173). His work clarifies the essentiality of studying test fairness, while allowing various interpretations across multiple areas of the testing process in a specific setting. His framework offers insights concerning fairness investigations that are directly connected to steps in testing activities of the testing process. Second, although social,
cultural, and political considerations play an important role in fairness investigations, an analysis that focuses on the socio-cultural context must examine various contextual forces that often compete with each other (Moss et al., 2008). Such examination of the social context requires a wider disciplinary collaboration with expertise in different academic backgrounds such as sociology, political science, and education (Moss et al., 2005). While the consideration of the socio-cultural forces in a specific context is important, it is not the focus of this research. Thus, there will be an underlying acceptance that the socio-cultural context influences test fairness, but the concentration of this dissertation will be on the testing process including test design and development, administration, scoring, and score-based use.

Willingham describes the conceptualization of test fairness in two works: his fairness framework that is centred around comparability for groups and individuals of test takers (Willingham & Cole, 1997), and his Fairness Manifold which adds the process, participants, and situation in addition to comparability (Willingham, 1999). In the following section, I elaborate on these four elements: comparability, process, participants, and situation. First, Willingham states that comparability for groups and individuals of test takers is the central principle in evaluating the fairness of a test (Willingham & Cole, 1997; Willingham, 1999). In Willingham’s work, comparability, as the organizing principle of test fairness, implies the comparison of different treatments of test takers, both groups and individuals. Unlike comparability used in comparability across different forms (discussed elsewhere in this chapter), Willingham’s definition of comparability has a broad meaning. Comparable treatment means “somewhat different things at different stages in the assessment process” (Willingham & Cole, 1997, p. 350), and it is subject to
different interpretations in different areas of the testing process that address “objective and impartial,” “unbiased,” and “equitable and just” considerations (p. 6). Comparability is used to represent multiple dimensions such as absence of bias, comparability across different forms, equality, and equity as discussed earlier. The use of the word *comparability* in defining test fairness is particularly helpful since it points out the key element of test fairness—comparing treatments between person and person and group and group, while at the same time allowing multiple dimensions across the testing process (e.g., equality, equity, and lack of bias). Clearly a test cannot be comparable in every aspect for every group and every test taker, but the goal is to make the test as comparable as possible for groups and individuals of test takers. For any empirical study of fairness issues, it is important to examine comparable treatment for both groups and individuals of test takers.

Second, in Willingham’s work, test fairness is examined within the testing process, which includes multiple aspects from test design and development, administration, scoring, to score-based use. Willingham’s work focuses on investigating test properties and examines activities and practices directly related to testing. As stated, Willingham and Cole’s view (1997) is echoed by the work of Xi (2010) and McNamara and Ryan (2011). A majority of the standards, guidelines, and codes of practices also suggest that fairness is about whether test takers are provided with opportunity to perform at their best in terms of, for example, test design and development (e.g., DIF and fairness review), administration (e.g., test standardization and accommodations), scoring (e.g., standard scoring procedures), or score-based test use (e.g., decision-making). This
element emphasizes the essentiality of studying test fairness across multiple areas of the testing process directly related to testing activities, not beyond.

Third, Willingham (1999) notes that test participants should be incorporated into the discussion of test fairness through “desegregating the pool of examinees or the various groups of professionals and institutions who have different connections with the process” (p. 224). He briefly explains that the views from test takers as well as professionals and institutions may conflict, but they all have a legitimate claim to fairness. Although Willingham does not provide clear suggestions concerning how different participants may be involved in fairness discussions methodologically, his position to legitimize stakeholders’ perceptions provides implications and directions. There is a need, in particular, for stakeholders from outside the technical community (e.g., test users and test takers) to discuss the fairness of testing policies and procedures that are often defined by policymakers and test developers (Madaus, 1994). The legitimacy of a test requires recognized acceptance and support among test stakeholders with solid empirical evidence. This element points out the importance of stakeholder involvement for fairness research.

Finally, when considering test fairness in his *Test Fairness Manifold* (1999), Willingham recognizes the role of the situation. By situation, he refers to a specific selection system. Willingham briefly contends that fairness within the testing process may have different meanings in different socio-cultural contexts, in particular, with regard to score-based decision making. Willingham argues that different selection models yield contradictory results for test takers; there has been no agreement on a preferred model. Testing policies and practices that seem to be fair in one context may appear
unfair in another context. Willingham’s view resonates with the legal debates (Cumming, 2007; US Commission on Civil Rights, 2002), and the reality that admission policies and practices vary substantively in different universities in the United States and around the world (Gipps & Stobart, 2009; Brelan, 1995, 2002). How the fairness of a test, such as the GSEEE, is influenced by its testing situation remains under-researched.

**Operational Model**

A variety of specific aspects across the testing process have been identified for fairness investigations in Willingham’s work (Willingham & Cole, 1997; Willingham, 1999). For instance, there are three major aspects in test development: sensitivity review, DIF, and comparability across different test forms. However, since their framework is mainly written for educational researchers and the technical community, other stakeholders such as test users and test takers are not involved in all the aspects of the testing process and they do not always have experience. Adaptation, hence, is needed in using this framework within the context of the GSEEE.

Thus, I have built an operational model to guide my dissertation research (see Figure 2). The operational model, which is designed specifically for this dissertation, provides a visual representation of the conceptual framework. The operational model, from the inner to outer circle, includes: comparability for groups and individuals of test takers (*comparability*), six aspects within the four areas of the testing process (*process*), three groups of test stakeholders (*participants*), and the GSEEE with arrows indicating dynamics of value judgments which are embedded in the testing situation (*situation*). Note that the three groups of stakeholders do not necessarily correspond only to the areas that they fall within. For example, the test user group often provides meaningful
information in terms of the fairness of score-based test use, as shown in Figure 2. However, other groups, for example, test takers, may also address issues of fairness in test use.

Figure 2. An Operational Model for the GSEEE

The operational model includes six specific aspects across the four areas of the GSEEE process. The choice of the six aspects is consistent with the literature review, and they are the most commonly investigated aspects, examining fair treatment for individuals and/or groups of test takers. Guided by the fairness literature, some of the aspects focus on comparability investigations for groups (1.1 test bias, 2.2 test accommodations, and 4.1 predictive bias), and some for individuals (2.1 standardized administration, 3.1 scoring procedure). The aspect of fairness in decision-making (4.2) is
examined for both groups and individuals of test takers. In the following section, I elaborate on the six detailed aspects.

Area 1: Design and development

1.1 Test bias: this form of investigation examines if groups of GSEEE test takers (e.g., gender and academic background) obtain comparable opportunities in terms of test construct, format, and content on a given test administration, so that they can demonstrate their knowledge and skills that are relevant to the purpose of the GSEEE.

Area 2: Administration

2.1 Standardized administration: this form of investigation examines if individual GSEEE test takers without disabilities take the GSEEE under uniform physical conditions and consistent and secure administration.

2.2 Test accommodations: this form of investigation examines if GSEEE test takers with disabilities and special needs are provided with accommodations so that their relevant knowledge and skills can be accurately demonstrated.

Area 3: Scoring

3.1 Scoring procedure: This form of investigation examines if performance of individual GSEEE test takers are accurately and consistently scored to demonstrate relevant knowledge and skills that they have acquired.

Area 4: Score-based use

4.1 Predictive bias: This type of investigation examines if the GSEEE is an effective and accurate predictor of subsequent learning for groups of test takers (e.g., gender and academic background).
4.2 Decision-making: This type of investigation examines if decisions with groups and individuals of the GSEEE test takers based on the GSEEE test scores along with other relevant information are made appropriately.

Specifically, in Chapter 3, I focus on 1.1 test bias and use Differential Item Functioning (DIF) and content analysis to investigate potentially biased test items towards different test taker groups in one GSEEE administration. Then, in Chapter 4, I examine test users’ and takers’ perceptions on the fairness of the GSEEE in all the six aspects of the testing process as listed in the model, including 1.1 test bias, 2.1 standardized administration, 2.2 test accommodations, 3.1 scoring procedure, 4.1 predictive bias, and 4.2 decision-making. Given the nature of the manuscript-method dissertation, Chapters 3 and 4 are written as stand-alone documents for journal publication. Combined, this doctoral dissertation intends to provide empirical information regarding the fairness of the GSEEE from the psychometric and stakeholder perspectives.
CHAPTER THREE

DIF INVESTIGATIONS ACROSS GROUPS OF GENDER AND ACADEMIC BACKGROUND IN A LARGE-SCALE HIGH-STAKES TEST

Introduction

High-stakes tests are a predominate tool to measure test takers’ knowledge and skills for the purpose of admission or certification. These tests are used as a means to classify, select, and judge individuals. Given the important consequences of high-stakes tests, there are concerted efforts to ensure that tests are fair to test takers. One example in obtaining empirical evidence of test fairness is to detect bias in the test in favour of or against test takers from certain groups (e.g., gender, linguistic, or ethnic status) that result in construct irrelevant differences in test scores (Cole & Zieky, 2001; McNamara & Roever, 2006). Differential item functioning (DIF) has become one of the most commonly used methods to judge whether test items function in the same manner for different groups of test takers. A similar procedure, differential bundle\(^8\) functioning (DBF), provides a measure of performance differences across clusters of items, typically grouped by “some organizing principles” (Douglas, Roussos, & Stout, 1996, p. 466). Although DIF and DBF are not sufficient to identify bias (Angoff, 1993; McNamara & Roever, 2007), they are valuable tools to explore irrelevant factors that might interfere with testing scores, discriminate against certain subgroups, and produce inaccurate inferences.

---

\(^8\) The term bundle refers to “any set of items choose according to some organizing principle” (Douglas, Roussos, & Stout, 1996, p. 466). Gierl (2005) described four general organizing principles: content, psychological characteristics (e.g., problem-solving strategies), test specifications, and empirical outcomes.
Over the recent decades, DIF research has been conducted with various second language tests (for reviews of DIF research in language testing, see Ferne & Rupp, 2007; Kunnan, 2000). These studies examined the effects of a variety of grouping variables such as gender (Aryadoust, Goh, & Kim, 2011; Breland & Lee, 2007; Takala & Kaftandjieva, 2000), language background (Elder, 1996; Kim & Jang, 2009), ethnicity (Freedle, 2006; Snetzler & Qualls, 2000; Taylor & Lee, 2011), age (Geranpayeh & Kunnan, 2007), and academic background (Pae, 2004) on language performance. As Ferne and Rupp pointed out (2007), DIF research in second language testing has “only just begun to investigate DIF for language tests outside North America” (p. 144). As such, this study examined DIF and potential bias with one of the large-scale high-stakes language tests in China—the Graduate School Entrance English Examination (GSEEE). In spite of the large number of the GSEEE test takers every year, DIF detection has not been conducted (He, personal communication, April 15, 2010). DIF investigations may present a relatively novel perspective to understand issues and concerns related to the fairness of the GSEEE.

The GSEEE is designed and administered by the National Education Examinations Authority (NEEA) of the Ministry of Education of the People’s Republic of China (Liu, 2010). The major purposes of the GSEEE are to measure English proficiency of test takers and to provide information for educational institutions to select candidates for their master’s programs (He, 2010). According to its test specifications, the GSEEE examines test takers’ linguistic knowledge in grammar and vocabulary, and skills in reading and writing (NEEA, 2009). The total number of test takers for the GSEEE administration, for example, in 2011, reached approximately 1.51 million and the
acceptance rate to enter master’s programs was 32.75% (MOE, 2011). This test has significant consequences for millions of test takers who compete for graduate program admissions and educational opportunities.

Considering the demographic information of China and target test-taking population of the GSEEE, the study investigated DIF and DBF across groups based on gender (female or male) and academic background (Humanities/Social Sciences or Sciences). First, a gender gap remains in the Chinese education and employment market (Postiglione, 2006). Females are less likely to receive higher education (Liu & Li, 2010). Although gender differences on language performance have long been examined in the literature (Cole, 1997; Kong, 2009), controversies exist regarding the interactions between gender and language performance, and it is unclear whether gender performance differences are due to test bias or ability differences. This study examined whether the GSEEE functioned differentially towards different gender groups and may bring advantages or disadvantages towards a certain gender group, which, as a result, could lead to a gender gap in educational and employment opportunities. Second, the GSEEE is designed for all non-English major test takers in any areas of Humanities, Social Sciences, and Sciences (He, 2010). Some literature has shown that test takers’ academic background knowledge facilitates language performance (Kintsch, 1998; Tapiero, 2007). It is possible that the GSEEE differentially and unfairly favours test takers from certain academic background, which affects their opportunity to obtain master’s education.

Examining how gender and academic background impact GSEEE performance has important implications to explore whether test taker groups are provided with equal opportunities to perform across a large country like China. Since tests may not always be
designed to keep the diversity of learner characteristics in mind, it is essential for test developers to monitor the test and examine its quality to see whether the test is fair to test taker groups (Geranpayeh & Kunnan, 2007). Specifically, the study addressed the following two research questions:

(1) Do the GSEEE items and bundles exhibit differential functioning toward test taker groups of gender (female or male) and academic background (Humanities/Social Sciences or Sciences)?

(2) If the GSEEE items and bundles exhibit differential functioning, what do test reviewers perceive are the possible causes? Can these causes be linked to potential bias toward test taker groups of gender and academic background?

Differential Item Functioning

Differential item functioning (DIF) is a statistical method to explore whether groups of test takers with equal ability have differing response probabilities of either successfully answering an item (i.e., in multiple choice) or receiving the same item score (i.e., in performance assessment) (Zumbo, 2007). According to Roussos and Stout (2004), 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” (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 occurs 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. In this study, the traditional, exploratory DIF approach was adopted. Although it may be preferable to conduct DIF analyses based on substantive, priori hypotheses using the
confirmatory approach, exploratory-based DIF analyses are still common in the test development and evaluation process (Walker, 2011). Using an exploratory DIF analysis paradigm may often be needed in practical evaluation of a test such as the GSEE. The traditional, exploratory approach has been used in previous empirical studies, using various DIF techniques (Pae, 2012; Woo & Dragan, 2012).

The exploratory approach is often conducted in two steps: statistical identification of items that favour a particular group followed by a substantive review of potentially biased items to locate the sources of DIF (Gierl, 2005). To conduct the first step, a number of statistical procedures have been developed and tested such as the Mantel-Haenszel method (MH), logistic regression (LR), and IRT (see a review by Clauser & Mazor, 1998). This study used the technique of Simultaneous Item Bias test (SIBTEST), considering the research questions. Developed by Shealy and Stout (1993), 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., male and female). 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. SIBTEST has been proven to be a powerful DIF procedure (Narayanan, 1994; Penfield & Lam, 2000; Zheng, Gierl, & Cui, 2007). It uses a regression estimate of the true score based on iterative purification instead of an observed score, which increases the accuracy of the matching variable. SIBTEST examines both uniform and non-uniform DIF. More importantly, SIBTEST is one of a few procedures that can evaluate bundle DIF, alternatively, DBF. Items with
small but systematic DIF may very often go statistically unnoticed, but when combined at
the bundle level, DBF may be detected (Douglas, Roussos, & Stout, 1996; Roznowski &
Reith, 1999; Takala & Kaftandjieva, 2000). In the GSEEE, all test items are embedded in
texts, and test takers answer questions based on their understanding of those texts
(bundles). As such, SIBTEST appears to be a useful tool to detect DIF and DBF. The
SIBTEST procedure determines the extent of the DIF, and classifies items as having
either negligible (A-level, |β| < .059) DIF, moderate (B-level, .060 < |β| < .087) DIF, or
large (C-level, |β| > .088).

The substantive analysis is then conducted after the statistical DIF analysis. While
DIF analyses identify differential performance across items or bundles, substantive
analyses determine the likely causes of the DIF and whether these causes are connected
with bias. The substantive analysis usually involves item reviews by subject-area experts
(e.g., content specialists or item writers) in an attempt to interpret the factors that may
contribute to differential performance between specific groups of test takers (Douglas,
Roussos, & Stout, 1996). 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. Substantial studies used content
analysis to identify potential bias, 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 &
Kunnan, 2007; Uiterwijk & Vallen, 2005). The judgment of the potential bias may not be
decisive; however, the primary focus is that, when it is possible, such judgments should
be supported by empirical data.
Gender and Test Performance

Gender differences in cognition and learning have long been examined (Dennon, 1982; Hamilton, 2008). Numerous early studies investigated gender differences in language proficiency performance, especially in terms of language skills, content, and test format/response types. The findings vary considerably with respect to language skills and ability, from conclusions that “girls have greater verbal ability” (Cole, 1997, p. 11) to “there are no gender differences in verbal ability” (Hyde & Lynn, 1988, p. 62) to “women obtained lower means than men on the verbal scale” (Lynn & Dai, 1993, p. 462). In terms of test content and topic familiarity, males appear to have an advantage on physical, earth, and space science items in language tests (Brantmeier, 2003). Studies focusing on item format effect generally concluded that multiple-choice (MC) items seem to favour males and open-ended items such as essay tend to favour females (Bolger & Kellaghan, 1990).

DIF methods provide an ideal way to examine gender effects on second language testing performance (Breland & Lee, 2007; Camilli & Shepard, 1993; Pomplun & Sundbye, 1999). Carlton and Harris (1992) examined gender DIF on the SAT. They found that overall reading comprehension was easier for the female group than the matched group of males, and males tended to perform better on antonyms and analogies than equally able females. O’Neill, McPeek, and Wild (1993) also extensively studied gender DIF across three testing forms of the GMAT. Their study reported that reading comprehension items were differentially easier for males than females matched on verbal ability, which seems to be contradictory to previous findings of Carlton and Harris (1992). Takala and Kaftandjieva (2000) examined gender differences with a small sample
on a Vocabulary Test in Finland. Although there were test items that seemed to exhibit DIF in favour of either females or males, the test as a whole was not gender-biased since the observed differences in the test results remained even after excluding the DIF items. The number and magnitude of DIF items favouring females was almost equal to those favouring males, cancelling the effect of the DIF items. DIF cancellation has also been found and discussed in other studies (Nandakumar, 1993; Roznowski & Reith, 1999).

A few gender DIF studies have been conducted using large-scale test data in China. Lin and Wu (2003) examined gender differences on the English Proficiency Test (EPT). Using SIBTEST for DIF analyses and DIMTEST⁹ for dimensionality investigation, they concluded that although the EPT did not demonstrate much gender DIF at the item level (2 items with C-level DIF and 11 with B-level DIF), DIF analysis with bundles of items provided evidence for a female advantage in listening, and a male advantage in cloze, and grammar and vocabulary. Kong (2009) analyzed gender differences in the reading comprehension section of the Test for English Majors--Band 4 (TEM-4). Based on a SIBTEST analysis, two items with C-level DIF favouring females and one item with B-level favouring females were identified. Two passages showed C-level DIF at the bundle level, with one favouring females and one favouring males. Expert review of DIF and DBF concluded that the potential reasons for the DIF existence might be related with gender topics and problem-solving items. However, as these gender topics were “not beyond the requirement of test specifications” (p. 17), Kong concluded that no test bias existed. Lastly, Lei (2007) examined the National Maculation English

---

⁹ DIMTEST examines the dimensional structure of a dataset and provides information about multidimensionality. However, caution is needed when using DIMTEST since DIMTEST results in a moderate to severe loss of power in some situations (Seraphine, 2000; Walker, Azen, & Schmitt, 2006).
Test (NMET) and did not find any gender DIF across the 90 multiple-choice items and the essay item. Hence Lei concluded that the overall gender differences on the NMET were due to real differences in English language abilities between males and females.

The above review of literature indicated that variation exists in data about the relationships between gender and language performance. This may be partially due to the fact that these studies investigated gender performance differences on tests that focused on various language skills and used different test format/responses types. These DIF studies helped to explore how tested skills, content, and format in a specific test might impact gender performance differences. DIF research is needed to investigate the gender effects on the large-scale high-stakes GSEEE, which has not been examined previously.

Academic Background and Test Performance

The interactions between test takers’ background knowledge and language proficiency, reading comprehension in particular, have been thoroughly studied in first language tests. Research papers consistently identify a facilitating effect of background knowledge on cognitive learning and reading comprehension theoretically and empirically (Kintsch, 1998; McNamara et al., 1996). The theory of Situation Models (Kintsch, 1998) describes how readers supplement the information provided by a text from their knowledge and experience in long-term memory to achieve a personal interpretation of the text that is related to this information in the text. First, in order to identify connections and effectively comprehend the text, the reader needs to process the text that consists of elements and relations within the text. Then, the reader must add nodes and establish links between nodes from his or her own knowledge and experience to make the structure coherent, complete gaps in comprehension, interpret information,
and integrate all information with prior knowledge. Successful comprehension requires not only an adequate processing of the language (the semantics of words), but also the reader’s familiarity with the situation described in the text that is gained through his or her interactions with the world and previous experiences. Needless to say, readers with abundant experience and domain knowledge tend to understand texts better than readers with little experience and domain knowledge (Kintsch, 1998; Tapiero, 2007).

In the second language area, some studies have been conducted to examine the role of academic background on test performance (Hill & Liu, 2012; Jennings, Fox, Graves, Shohamy, 1999). Studies suggest there is a relationship between subject area knowledge and test performance (Chung & Berry, 2000; Krekeler, 2006). Hale (1988) examined if students’ academic disciplines interacted with the text content in determining performance on the reading passages of TOEFL. Results found that students in the Humanities/Social Sciences and Biological/Physical Sciences performed better on passages related to their own background knowledge than on other passages. The effect was significant for three of the four test forms; however, Hale concluded the effect was relatively small since the apparent subgroup advantage translated into “only a few points in overall TOEFL scale score” (p. 59). Hale attributed this to the TOEFL reading passages that were drawn from general readings rather than specialized textbooks. Examining the effect of background knowledge on the IELTS with test takers with a range of academic levels, Clapham (1996) also found that students generally performed significantly better on the reading modules in their own subject areas.

Despite the identified links between academic background and language proficiency, few researchers have used DIF methods to explore such interactions. Pae
(2004) used the MH procedure and the IRT likelihood ratio approach to investigate test performance between test takers in the Humanities and Sciences. The study found that seven items favoured the Humanities test takers, and 9 favoured the Science test takers. Although there might exist some cancellation effects, the score differences due to the influence of potential bias were not considered as negligible. The preliminary content analysis of the test indicated that items dealing with science-related topics, data analysis, and number counting were differentially easier for the Sciences, whereas items about human relationships were differentially easier for the Humanities. It is unknown whether and how test items of the large-scale high-stakes GSEE demonstrate DIF and potential bias toward different academic groups.

Method

This section described subjects, the GSEE administered in 2009, and data collection and analysis procedures. Before the study was conducted, clearance from the Queen’s University General Research Ethics Board had been received (see Appendix A for clearance letter). In addition, prior to content analysis, all the test reviewers were given a letter of information that detailed their involvement in the study, and they signed a consent form (see Appendix C & D).

Subjects

The study used the GSEE item-level data of the 2009 administration from one major university in South China. This university provides a wide range of master’s programs; hence thousands of applicants across China are attracted to apply for these master’s programs offered by the university every year. In this study, applicants’ background information and their GSEE 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). In addition, 25.4% of the test takers had graduated in previous years and 74.6% were in the last year of their undergraduate programs. These pieces of information were similar to the demographic information of the overall GSEEE test-taking population (MOE, 2009).

*The Graduate School Entrance English Examination*

The 2009 administration of the GSEEE consisted of three sections (See Table 2). Section I, Cloze, was a multiple-choice (MC) test of vocabulary and grammar with 20 blanks in one text. 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. According to the test specifications, a practical writing task refers to a writing task used in everyday situations such as personal and business mail, memos, and outlines. Whereas, an essay writing task requires test takers to produce a written argument or discussion on a given topic and give some examples to support their

---

10 The GSEEE administered in 2009 consistently used the term text which consisted of more than one passage. This study used the same term throughout this paper.
points. Test takers had to write about 100 words in the first task and 200 words in the second.

Table 2

*Description of the GSEEE Administered in 2009*

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

The Section I Cloze (Items 1-20) and Parts A and B in Section II (Items 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 these three criteria was penalized 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 that addresses all major elements of the task, demonstrates syntactic variety and range of vocabulary, uses accurate word choice and proper grammar, reveals clear coherence and smooth progression of ideas, and displays appropriate choices of forms and registry.

Data Analyses

Descriptive statistics were first 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 provide an estimate of the internal consistency of the GSEEE. After that, the two-step exploratory approach was conducted (Gierl, 2005). In this study, SIBTEST was first used to identify the presence of DIF and DBF, followed by the content analysis that explored the likely causes of DIF and DBF towards the test taker groups of gender and academic background.

SIBTEST

The study used female test takers and test takers from Humanities/Social Sciences as the focal group and male and Sciences 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, Roussos, & Stout, 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 order to guard against unrepresentativeness within each group of gender and academic background, I used an equal number of test takers with different characteristics to facilitate comparisons. In other words, when examining gender effects on the GSEEE, a stratified sample of 1000 female test takers from Humanities/Social Sciences and 1000 female test takers from Sciences was selected as the focal group; and a sample of 1000 male test takers from Humanities/Social Sciences and 1000 male test takers from Sciences was selected as the reference group. The sampling method was also applied to the investigation of academic background effects on the GSEEE. 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 possible causes of DIF and DBF. It also examined whether test reviewers perceived that these possible causes were linked to the potential bias toward groups of gender and academic background. The expectation of was that if test reviewers thought the differential functioning of these flagged items/texts
was due to variables that were irrelevant to the construct measured by the GSEEE, then it might be possible to determine if such items were biased. To complete the content analysis, recorded telephone interviews were conducted with the three reviewers. The reviewers were purposely chosen based on their gender, age, work experience, and extensive knowledge of English teaching and testing. They were current university professors with extensive teaching experience in both undergraduate and graduate programs (see Table 3). They all received professional training in Applied Linguistics with an emphasis on language testing and assessment. All of the test reviewers had participated in the test design of large-scale high-stakes English tests in China, and two of them were involved in GSEEE item writing. Since individual reviewers with various backgrounds could be expected to interpret the sources of each DIF/DBF in different ways arguably, this would result in a more comprehensive understanding of these flagged test items/texts.

Table 3

Background Information of Content Reviewers

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Age</td>
<td>46-50</td>
<td>40-45</td>
<td>51-55</td>
</tr>
<tr>
<td>Education</td>
<td>PhD</td>
<td>PhD</td>
<td>M. A.</td>
</tr>
<tr>
<td>Professional experience</td>
<td>25 years of teaching experience in English</td>
<td>more than 20 years’ teaching experience in English</td>
<td>24 years of teaching experience in English</td>
</tr>
<tr>
<td></td>
<td>Involvement in high-stakes item writing</td>
<td>Involvement in high-stakes item writing</td>
<td>Involvement in high-stakes item writing</td>
</tr>
<tr>
<td></td>
<td>Language testing researcher</td>
<td>Language testing researcher</td>
<td>ESL researcher</td>
</tr>
</tbody>
</table>
The format of the content analysis (see Appendix E) was similar to that conducted by Geranpayeh and Kunnan (2007). First, the three participants were asked to decide whether the flagged items/texts were likely to advantage/disadvantage test takers who were female or male and from Humanities and Social Sciences or Sciences background. They were asked to consider various sources of potential bias including context, contents, semantics, vocabulary, pragmatics, or any other potential sources. 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

Descriptive Statistics and Test Evaluation

Table 4 reports the mean scores, standard deviation, skewness, and kurtosis for each group and overall. The descriptive statistics showed that female test takers outperformed males and test takers from Humanities and Social Sciences outperformed the ones from Sciences based on the total mean scores. Skewness and kurtosis values ranged between +1 and −1, indicating that the distribution of the data could be considered normal. Cronbach’s alpha with each section and the total scores were calculated. In general, these reliability estimates were not high (α =0.53 for Section I; α =0.61 for Section II; α =0.65 for Section III; and α =0.71 for total).
Due to the low coefficient estimates, a follow-up investigation was conducted to examine item quality by using IRT-Bilog index (See Appendix F). Generally speaking, the test showed a wide span of item difficulty with $P$-values (proportion correct) ranging from .09 to .85. However, item discrimination values based on the point-biserial Pearson correlations were generally low, ranging from .02 to .35, with a large number below .20 (29 out of 45 MC items). In addition, two items—Item 12 and 43 had negative item discrimination values (-.07 and -.04 respectively). These values show that the GSEEE test items did not function well to differentiate the high performers from low performers.

**SIBTEST Results**

Table 5 provides an overall description of the SIBTEST results at the item and bundle (text) level. To identify if the item quality may have an impact on the DIF results, the DIF/DBF analysis was conducted with and without the two test items that showed negative discrimination values. Results found that the quantity and size of the flagged items and bundles remained even after excluding these two items (see Table 5).
Table 5

Results of the SIBTEST analysis

<table>
<thead>
<tr>
<th>Grouping variable</th>
<th>Section</th>
<th>Item/Bundle</th>
<th>Beta Uni with/without</th>
<th>P-value</th>
<th>Favouring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>I Cloze</td>
<td>Item 14</td>
<td>-.075/-0.075</td>
<td>0.00</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text (Items 1-20)</td>
<td>-.326/-3.15</td>
<td>0.00</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>II RC</td>
<td>Part A Text 2</td>
<td>.082/.081</td>
<td>0.03</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Items 26-30)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Part A Item 37</td>
<td>.081/.081</td>
<td>0.00</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>III Writing</td>
<td>Essay writing</td>
<td>-.251/-2.43</td>
<td>0.00</td>
<td>Female</td>
</tr>
<tr>
<td>Academic background</td>
<td>II RC</td>
<td>Part A Item 30</td>
<td>.160/.159</td>
<td>0.00</td>
<td>Sciences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Part A Text 2</td>
<td>.160/.160</td>
<td>0.00</td>
<td>Sciences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Items 26-30)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

Regarding gender effects on the GSEEE, the SIBTEST analysis at the item level indicated that Item 14 in Section I showed B-level DIF favouring females while Item 37 in Section II Part A showed B-level DIF favouring males. Also, the essay writing task in Section III Part B showed C-level DIF favouring female test takers. Specifically, Item 14 embedded in the text animal intelligence passage examines logical relationships; test takers were required to select the best word from four choices: (A) by chance (B) in contrast (C) as usual (D) for instance. Item 37 asks the test takers to determine the inferencing idea from one of the paragraphs in the text regarding the history of the New World. The essay writing task asks test takers to write an essay about closeness and remoteness of Internet based on a picture. With respect to the bundle DIF (DBF), results showed that the Cloze text regarding animal intelligence in Section I favoured females while Part B Text 2 regarding genetic testing in Section II favoured males.
In terms of academic background effects on the GSEEE, the SIBTEST analysis at the item level found only one item, Item 30, which favoured the Sciences test takers. Item 30 asks test takers to identify an appropriate title for the text regarding genetic testing. When the bundle DIF came into play, Text 2 regarding genetic testing in Section II Part A showed significant DIF favouring the Sciences test takers.

Content Review

Three test reviewers examined what caused DIF/DBF with these flagged items/texts and whether these items/texts showed bias towards test taker groups of gender and academic background. Table 6 presents the ratings of the reviewers on the flagged items/texts.

Table 6

<table>
<thead>
<tr>
<th>Grouping variable</th>
<th>BIF Item/Bundle</th>
<th>Content analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reviewer A</td>
</tr>
<tr>
<td>Gender</td>
<td>Item 14</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Section I Cloze (Items 1-20)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Section II Part A</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Text 2 (Items 26-30)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Part A Item 37</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Section III Essay writing</td>
<td>3</td>
</tr>
<tr>
<td>Academic background</td>
<td>Section II Part A Item 30</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Section II Part A</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Text 2 (Items 26-30)</td>
<td>4</td>
</tr>
</tbody>
</table>

Gender DIF and DBF

Item 14 was of average difficulty (62% correct) with good discrimination (0.40). None of the three test reviewers concluded Item 14 showing bias towards male/female
test takers. Item 14, which was embedded in the Cloze, examined inter- and intra-substantial logical relationships. The answer for this item was “for instance,” and this phrase was listed in the glossary of 5500 words that were required for the GSEEE test takers (NEEA, 2009). The three test reviewers provided stereotypical comments and they thought that the differential functioning of Item 14 between males and females was due to learner characteristic differences that females tended to pay much attention to details and often revisited the sentences/questions.

None of the three test reviewers rated the Cloze section (Items 1-20) about animal intelligence as exhibiting an advantage or disadvantage towards female or male test takers from its content, format, plot, or test takers’ knowledge. Based on their teaching experience, the reviewers concluded that females generally outperformed males in identifying detailed information and noticing subtle changes in language. In comparison, males had the tendency to rush through the cloze items, often missing essential details that were necessary for a correct response.

Section II Part A Text 2 (Items 26-30) discussed the genetic testing of DNA. There was a gap in the conclusion of the three reviewers, and the three reviewers did not agree whether potential bias existed. Test Reviewer B felt that males performed better than their peers mainly because of males’ advantages in science topic and content. Hence, this text may be slightly biased and favour males. In comparison, Test Reviewer A and C did not conclude any bias towards female or male test takers. They claimed that discrepancies in English language proficiency were the major reason that caused the differential functioning of this text towards gender groups.
Item 37 was one of the more difficult items (42%) in the 2009 GSEEE with the low discrimination (0.21). None of the three test reviewers rated this item as biased towards female or male test takers. The item was embedded in Text 4 discussing the history of the New World. The primary target of this item was to correctly identify the main idea of one paragraph. The reviewers offered clichéd comments, suspecting that the existence of DIF was due to the situation that females generally lacked knowledge and interest in history.

Regarding Section III the essay writing task that required test takers to write an essay regarding the Internet based on a drawing, none of the three test reviewers rated the DIF existence as biased towards female or male test takers. The test reviewers pointed out that the differential functioning might be due to the situation that females were generally better than males in productive skills of speaking and writing. The reviewers also mentioned that female test takers compared with males were more motivated and diligent, and more likely to take time on memorizing words and sample essays.

**Academic background DIF and DBF**

Item 30 was not hard (71% correct), but had a low discrimination index (0.11). The test reviewers had conflicting views about whether this item showed bias toward test takers with different academic backgrounds. Item 30 required test takers to identify an appropriate title for the text regarding genetic testing. Test Reviewer B concluded that Sciences test takers would benefit from their background knowledge slightly since they were generally more familiar with this topic. In contrast, the other two test reviewers, Test Reviewer A and C, concluded no bias towards test takers from different academic backgrounds. They stated that Item 30 was a global question since it asked test takers to
generalize the title of the text, and it did not involve anything very complicated for test takers from the non-Sciences background.

Conflicting views among the reviewers also existed in terms of Section II Part A Text 2 (Item 26-30), which asked questions related to DNA testing and its problems. Test Reviewer C did not think this text showed bias towards test takers from different academic backgrounds since the discipline-related topics such as the DNA testing fell within the scope of the GSEEE test specifications. In contrast, Test Reviewer A and B rated the text as favouring test takers from the Sciences background. They felt that the whole text was a passage of English for Specific Purpose (ESP) reading, which was beyond the scope of the GSEEE test specifications. The text was more than just general knowledge, and it was likely to favour those who had content knowledge. Therefore, they concluded that this text was biased slightly favouring test takers from the Sciences background.

Discussion

This study investigated group membership effects on the GSEEE administered in 2009 regarding gender and academic background. Descriptive statistics found low reliability and discrimination values. The results also found two items with negative discrimination values, indicating that good performers tend to answer them wrong while poor performers get them right. Such results show the existence of flawed items in the GSEEE administrated in 2009. The GSEEE does not provide test takers with equal opportunities to demonstrate their knowledge and skills. The GSEEE may underestimate (or overestimate) test takers who could perform better (or worse) and undermine the fairness claim for individual test takers. Further, SIBTEST found the existence of DIF
and DBF towards the groups of gender and academic background. Based on the results of content analysis, three test reviewers identified a multitude of factors that made these items/texts easier or harder for different groups. In the following, I will discuss the findings based on SIBTEST and content analysis.

**SIBTEST Findings**

SIBTEST was used to explore the presence of DIF and DBF and quantified the size of DIF and DBF. In terms of gender groups, the current study identified two items/texts favouring males at B level and three items/texts favouring females at C level. The results seem to show no systematic relationship between the DIF direction and item difficulty/item discrimination values. The results regarding the flagged items/bundles favouring males might be partially related with test content and topic familiarity. As the previous literature indicated, males tended to perform better in “scientific-related content” than their matched female peers (Bond, 1993, p. 277). The study has also provided evidence that test format might be another reason to cause the differential functioning of the items/texts between male and female test takers. Compared with the practical writing task (10 out of 100 points), the effect of test format on gender performance differences is more evident in the essay writing since the essay writing accounted for a larger percentage of the total score (20%). In this study, females performed better than males in essay writing, which is consistent with the conclusions from previous studies (Bolger & Kellaghan, 1990; Pomplun & Sundbye, 1991). Moreover, female test takers performed significantly better than males on the Cloze. This result conflicts with Lin and Wu’s study (2003) in which the bundles of Cloze favoured males slightly. As Cloze is a commonly used testing format in language testing
(Bachman, 1990), how gender interacts with the testing performance on Cloze certainly merits further research in the future.

Regarding the group of academic background, SIBTEST found that one item and one text functioned differentially towards the Sciences test takers. These results might be related to the content of the text, which focused on specific knowledge of Genetic Testing. This is consistent with results from the previous empirical studies. Hale (1988) found students in Humanities/Social Sciences and Biological/Physical sciences performed better on passages related to their own background knowledge than on other passages. Pae (2004) concluded that items favouring the Sciences test takers were generally associated with science-related topics. A facilitating effect of background knowledge on reading comprehension is evident in this study.

**Content Analysis**

While the SIBTEST results indicate that interactions existed between gender, academic background, and the GSEEE test performance, the determination of test bias warrants further investigation through a content review of the test. Three experts in language testing examined the likely causes of the flagged items/texts. They examined, based on their perceptions, whether these causes were linked to bias. Regarding gender DIF, the reviewers believed there were multiple factors associated with performance discrepancies. These factors include general, often stereotypical comments that female test takers were advantaged in identifying information and subtle changes in language, in the productive skills of speaking and writing, and more highly motivated and diligent, while males were more knowledgeable in history and scientific-related topics. However, the three reviewers did not agree and it was not consistent whether these identified
reasons decisively led to bias among the three reviewers. The three reviewers claimed that, overall, gender played a minimal, negligible role in affecting test takers’ admission status for master’s programs in China.

On the whole, the reviewers provide only anecdotal comments, suggesting that these reviewers may not have gone beyond the notion of stereotyping input. The reviewers generally do not appear to provide in-depth thoughts about potential causes of the flagged items/texts. Their views are only at a superficial level, which may be partially due to their role in the GSEEE (as item writers) and fear of making testing and MOE look bad. Moreover, in spite of working as language experts in universities and being involved in high-stakes item writing, some reviewers seem unclear about what bias meant after having heard on the explicit explanation about the nature of the study. This indicates that it is of importance to provide testing guidelines and extensive training in the GSEEE test development agenda to keep a balanced selection of items and texts for the goal of test fairness. Additionally, the results of the content analysis reflect the challenge of using expert judges for bias analyses in DIF research, which is consistent with the previous studies (Geranpayeh & Kunnan, 2007; Uiterwijk & Vallen, 2005).

Similar stereotypical comments were found regarding the factors associated with performance discrepancies towards groups of academic background among the three reviewers. In addition, the reviewers had different views regarding whether the flagged item and text were linked to potential bias. Such discrepancy reflects two issues. First, the reviewers seem to disagree whether the flagged item and text was disciplinary specific, alternatively, whether the item/text examined general English or English for Academic Purposes (EAP)/English for Specific Purposes (ESP). To further understand the lexical
component of the disputable text— the Genetic Testing, a lexical text analysis was conducted using Vocabprofile (Cobb, 1999). Results found the passage did not seem very academic because of the low percentage of Academic Word List (AWL) words (5.5%). However, there were about 13% less frequent words from the Off-list category such as *paternity, prescription, kinship, genetic, saliva*, and *ancestor*. The top three key words for the proper comprehension of the text all fall into the scope of the Off-list category — *paternity, genetic, and ancestor*. Considering the Chinese national policies and institutional practices promoting the use of English as a teaching language in disciplinary courses (Du, 2003; Yang, 1996), test takers with Sciences background may have opportunities to learn these Off-list category words in their disciplinary courses. This suggests that test takers from Sciences background would be more advantaged than those from Humanities/Social Sciences background in this disputable text.

Second, the results of the content analysis with the two flagged items/texts towards the academic background groups also reflect the reviewers’ differing interpretations of the GSEEE test specifications. It remains a point of disagreement among the three reviewers whether the GSEEE is designed to test General English, EAP, or ESP. While one reviewer believed the discipline-related topics such as the DNA testing fell within the scope of the GSEEE test specifications, the other two reviewers stated that ESP readings were beyond the scope of the GSEEE test specifications. According to the 2009 GSEEE manual (NEEA, 2009):

The GSEEE is designed for non-English majors. Considering the practical purposes, test takers should have command of vocabularies related to one’s profession or academic major as well as those involved in individual likes and
dislikes, life habits, and religion… Test takers should be able to comprehend not only books and newspapers in a variety of topics, but also literature, technical manuals, and introduction of products related to one’s academic or professional area. (p. 2)

It would appear from the above description that the purposes of the GSEEE are to examine test takers’ knowledge and skills in General English as well as EAP and ESP. However, this may create confusion and difficulties in test design and development because a great part of academic and professional vocabulary and reading comprehension consists of vocabulary, genres, and discourse that are discipline-specific and context-situated. The differences among General English, EAP, and ESP have long been discussed in language teaching, learning, and testing (Dudley-Evans & St John, 1998; Fowerdew & Peacock, 2001). While General English focuses on day-to-day communication, EAP and ESP are to meet specific needs of learners—academic learning and professional developments (Dudley-Evans, 1997). Thus EAP and ESP are centered on the language appropriate to these activities in terms of grammar, lexis, register, discourse, genre, and study skills. How to select and produce more balanced test items and serve the multiple purposes in assessing test takers’ ability in General English as well as EAP and ESP presents a challenge for the GSEEE test developers.

Conclusion and Implications

Since high-stakes tests play a significant role in decision-making, it is important to examine how tests function towards groups of test takers and what they really measure. This study found the presence of DIF and DBF on the 2009 GSEEE towards groups of gender and academic background. A review of the flagged items and tests by the three
test reviewers identified myriad factors that potentially contributed to different performance of focal and reference group members who were matched on ability. Nevertheless, consistent evidence was not found in the content review to suggest these flagged items/texts exhibited bias.

While systematic bias may not have been detected, the study provides important implications for the GSEEE test practices and English education in China as a whole. First, since significant group differences exist due to passage content/topic familiarity and test format/response types, there are implications for curriculum developers to cover a variety of topics related to Humanities, Social Sciences, and Sciences in curriculum design and for classroom teachers to use different teaching methods to enhance students’ learning on certain topics and formats (e.g., Sciences topics for students from Humanities background). Second, there are some major implications for the GSEEE test developers and item writers. The purposes of the test need to be clearly described. When the test specifications are developed, it is important to seek balance (e.g., test format, content) and construct a test that is reasonably suited to test taker groups. Test developers need to consider what to test and how to test as well as how the choices of these issues may impact fairness for different groups. Proper training and item writing guidance should be provided to help item writers to be aware of fairness issues such as item bias. Third, more significantly, the study shows the urgency to improve the item quality of the GSEEE. 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 (Fu & Jin, 2012), the paper calls for moderation panels to conduct ongoing technical examinations and review draft test materials in a systematic manner.

Since this study is exploratory by nature, more in-depth, systematic inspections using confirmatory approaches are warranted in the future. The confirmatory approach for DIF analyses is theory-driven and allows for more thorough explanations of DIF (Ferne & Rupp, 2007; Sandilands et al., 2012). Rousos and Stout’s (1996) two-stage multidimensionality-based confirmative DIF approach could provide direction for future studies. In addition, the use of multiple DIF procedures, such as Logistic regression, IRT, or MH alongside with SIBTEST, will be helpful to cross-validate statistical results and increase the certainty in identifying flagged items. As the groups of academic background (Humanities & Social Sciences or Sciences) were actually not equally represented in the real world population in China, inflation may exist and caution is needed to interpret the DIF results. Given the complex nature of test fairness, the current study only focuses on one aspect of test fairness: item bias in test design and development. Besides creating the best possible test items to eliminate bias, it is imperative to examine test fairness when tests are administered, scored, and used, and to investigate fairness from multiple perspectives of test stakeholders such as test users and test takers.
CHAPTER FOUR
FAIRNESS IN A LARGE-SCALE HIGH-STAKES LANGUAGE TEST:
ADMINISTRATORS’, TEACHERS’, AND TEST TAKERS’ PERCEPTIONS

Introduction

Test fairness has been pursued for many years in high-stakes testing. Such pursuit can be traced back to the Chinese Imperial Examinations (606 AD-1905 AD) in which rigorous procedures were used to ensure fairness and equality, for example, double marking and test scripts being copied so that test takers’ handwriting would not be recognized (Cheng, 2010; Yu & Suen, 2005). Later, the 20th century witnessed the evolution of research in test fairness in examinations of statistical bias (Cleary, 1968; Holland, 1985). Today, investigations of fairness continue to examine bias and the measurement aspect of testing (Banks, 2013; Fischer, Schult, & Hell, 2013).

Simultaneously, there are more and more discussions regarding value judgments of the fairness of a test (Camilli, 2007; McNamara & Roever, 2006).

Value judgments entail understandings and interpretations that give meaning to utterances, perceptions, and opinions which test stakeholders express in their life. The predominant views associated with test fairness are usually those of policy makers and test developers, and are demonstrated through testing policies, guidelines, and practices. Articulations of test fairness from outside the technical community such as those of test users and takers are often under-addressed (Hamp-Lyons, 2000; Shohamy, 2001). The stakeholder approach in examining test fairness has been widely used in psychology and business, for example, test-taking for job applications (Bacdayan & Geddes, 2009;
However, empirical studies based on test users’ and takers’ perceptions guided by a fairness framework are limited in language testing. In this study, I explored stakeholders’ perceptions of the fairness of a large-scale high-stakes test in China — the Graduate School Entrance English Examination (GSEEE). The GSEEE is one of the measures of the Graduate School Entrance Examinations (GSEE) that is used to determine if test takers from non-English majors can gain annual admission into a master’s program in China. The GSEEE has two purposes: to measure English proficiency of test takers and to provide information for institutes to select candidates for their master’s programs (He, 2010). The GSEEE has important consequences for test stakeholders such as test takers, graduate program administrators, teachers and supervisors, test developers, parents, and ultimately society as a whole.

I examined perceptions of the two subgroups of test users—graduate program administrators and English teachers as well as the test taker group. These stakeholders are most significantly influenced by the GSEEE. Graduate program administrators are crucial to this study because they make critical decisions based on the GSEEE test scores. They play a key role in screening applicants and ensuring that selected candidates have adequate English proficiency skills to complete master’s programs. The quality of students they admit into master’s programs is often related to the evaluation of institutional effectiveness, the available funding provided by the Chinese Ministry of Education (MOE), teaching and learning progress, and school reputation. Furthermore, the perceptions of teachers who teach English courses at the master’s level are important in examining fairness issues. Teachers work with test takers in classroom teaching and
learning after the successful applicants enter master’s programs; the English proficiency levels of these test takers establish a baseline for classroom teaching\textsuperscript{11} in master’s programs. These teachers may also work closely with program administrators to decide if test takers can enter master’s programs.

In addition, the GSEEE has a direct, significant impact on test takers’ life options and career opportunities. Test takers’ scores on the GSEEE determine if they may have a chance to be admitted to master’s programs in China. Although many test takers write high-stakes tests like the GSEEE every year, chances are that “virtually none [test takers in any test, italics added] will have participated in the test’s design, in writing test items, in critiquing the test methods, in setting cut scores or in writing or commenting on the performance descriptions that tie to their all-important scores” (Hamp-Lyons, 2000, pp. 580-581). Therefore, it may be helpful to involve test takers and explore their perceptions about the fairness of the GSEEE.

Through examining administrators’, teachers’, and test takers’ perceptions about the fairness of the GSEEE testing policies, guidelines, and practices, this study intends to empirically investigate whether the GSEEE test takers are provided with fair opportunities to perform. The study also explores these stakeholders’ overall perceptions of the fairness of the GSEEE as well as priorities and major aspects that influence their perceptions. The success of a large-scale high-stakes test is related to stakeholders’ perceived legitimacy and acceptance. Obtaining input from test users and takers has the potential to identify concerns, improve the test, and increase awareness, communication,

\textsuperscript{11} Based on the Syllabus for Non-English Major Master’s Students (1992), master’s students across China have to obtain certain credits on English courses at the master’s level to obtain a master’s degree, unless specified otherwise.
and transparency about the test. Note that the goal of this study is not to provide solutions to fairness concerns, but rather to offer empirical evidence regarding perceptions of the fairness of the GSEEE and provide avenues for further investigation, for instance, score-based research conducted by policy makers and test developers. Specifically, this study addressed the following three research questions:

1. What do the GSEEE stakeholders—graduate program administrators, English teachers, and test takers know about the GSEEE?
2. How do they perceive the fairness of the GSEEE testing policies, guidelines, and practices for groups and individuals of test takers?
3. What are their overall perceptions of the fairness of the GSEEE? What are the major considerations for such judgments?

The Graduate School Entrance English Examination

The GSEEE is one of the measures within the Graduate School Entrance Examinations (GSEE) that includes the first round of preliminary examinations and the second round of re-examinations. The Ministry of Education (MOE) makes operational decisions about the first round of preliminary examinations including time, place, content, and subjects. In the second round, these important decisions are determined by individual institutes. The first round of preliminary examinations includes four tests: Foreign Languages (English, Japanese, or Russian), Political Science, and two additional subject areas. The GSEEE is one of the Foreign Language tests, and a large majority of applicants choose English as their foreign language (He, 2010). Test takers whose scores surpass the cut-scores on all the four tests (including the GSEEE) and total scores may enter the second round of re-examinations.
The National Education Examinations Authority (NEEA), appointed by the MOE, is responsible for the design, development, and evaluation of the GSEEE (Liu, 2010). All the universities and research institutes across China, over 900 of them, use national cut-scores to decide if test takers may enter the second round of re-examinations, except for 34 universities\(^\text{12}\) whose master’s programs have the flexibility to set up their own cut-scores (commonly called independently determined cut-scores). There are a number of different national GSEEE cut-scores based on regions, ethnicities, and disciplines. For example, in 2012 (see Table 7), the whole country was divided into two broad region categories\(^\text{13}\) (Regions A and B) and 16 subcategories within each category based on ethnicity and disciplines (MOE, 2012). Region A included 21 of the more economically developed provinces while Region B included 10 of the less developed provinces including all five ethnic minority areas. In general, test takers from the ethnic majority group (commonly called the Han, about 92% of the total population) who applied for master’s programs located in Region A would have to pass the highest cut-score, followed by those who applied from programs in Region B. Test takers from the ethnic minorities (55 in total) would require the lowest cut-score. For example, in the discipline of Education, the highest cut-score was 40 points out of 100 or 40% for the Han majority who applied for programs in Region A, then 37% for those who applied for programs in Region B, and finally 30% for ethnic minorities.

---

\(^{12}\) The 34 universities, all top Chinese universities, have been granted the autonomy to decide their cut-scores since 2002.

\(^{13}\) Before 2012, there were three broad area categories. The current Region A used to include two regions: 18 more economically developed provinces and municipalities and three provinces and municipalities in north-western and south-western China.
Table 7

*The 2012 GSEEE National Cut-scores (MOE, 2012)*

<table>
<thead>
<tr>
<th>Ethnicity/Discipline</th>
<th>Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Region A</td>
</tr>
<tr>
<td>Ethnic minority</td>
<td>30%</td>
</tr>
<tr>
<td>Agronomy</td>
<td>33%</td>
</tr>
<tr>
<td>Physical Education</td>
<td>34%</td>
</tr>
<tr>
<td>Visual Art</td>
<td>34%</td>
</tr>
<tr>
<td>Chinese Traditional Medicine</td>
<td>34%</td>
</tr>
<tr>
<td>Philosophy</td>
<td>38%</td>
</tr>
<tr>
<td>History</td>
<td>38%</td>
</tr>
<tr>
<td>Science</td>
<td>38%</td>
</tr>
<tr>
<td>Engineering</td>
<td>38%</td>
</tr>
<tr>
<td>Medicine (Chinese Traditional Medicine excluded)</td>
<td>38%</td>
</tr>
<tr>
<td>Education (Physical education excluded)</td>
<td>40%</td>
</tr>
<tr>
<td>Military Science</td>
<td>41%</td>
</tr>
<tr>
<td>Law</td>
<td>42%</td>
</tr>
<tr>
<td>Management</td>
<td>50%</td>
</tr>
<tr>
<td>Economics</td>
<td>50%</td>
</tr>
<tr>
<td>Literature (Visual Art excluded)</td>
<td>52%</td>
</tr>
</tbody>
</table>

The setup of both the national and independently determined cut-scores is related to a quota system — Enrollment Planning Objectives (EPOs), which is stipulated annually by the MOE and provincial education committees (MOE, 2011; Wang, 2005). The EPOs are formulated, based on the national plans, teaching, research, and supervising capability of educational institutes, and the available financial resources (Wang, 2005). They determine how many spaces are available among master’s programs for the ethnic minorities and for the 15 major disciplines in a given year across the whole country. Alternatively, the number of master’s students that a university and its programs
can admit is approved by the MOE in advance. If the number of test takers who surpass the cut-scores in the first round is more than the admission quota for a particular program, a percentage will be used to decide who can enter the second round of re-examinations. The percentage between the quota and the number of test takers who can enter the second round usually ranges from 1:1.2 to 1:1.3 (Wang, 2005). For example, if a master’s program obtains a quota of 100 spaces from the MOE in a given year and 200 test takers pass the national or independently determined cut-scores, all the applicants will be ranked from highest to lowest and the cut-score will be set to allow test takers between No. 1 and No. 120 or No. 130 to enter the second round of re-examinations.

According to the GSEEE test specifications, the GSEEE examines test takers’ English language abilities, including knowledge of the language (grammatical competence, textual competence, and sociolinguistics competence) and skills in reading and writing (NEEA, 2012). Currently, the GSEEE assesses only reading and writing. Speaking has never been included, and listening was included only briefly between 2002 and 2005 due to administrative reasons (He, 2010). Speaking and listening are tested in the second round of re-examinations if test takers pass cut-scores in the first round. The GSEEE scores are valid for one year. The total number of the GSEE test takers, for example, in 2011, reached approximately 1.51 million, and the acceptance rate was 32.75% (MOE, 2011).

Conceptual Framework and Operational Model

Willingham’s fairness conceptualization, both his framework (Willingham & Cole, 1997) and his Test Fairness Manifold (Willingham, 1999), is used to guide this study. According to Willingham’s work, test fairness is rooted in a consideration of
comparability investigations for groups and individuals of test takers across the testing process from test design and development, administration, and scoring, to score-based use, involving stakeholder groups within the testing situation. His conceptualization consists of four elements: comparability, process, participants, and situation. Willingham’s framework highlights the essentiality of test fairness — comparable treatment between person and person and group and group, while simultaneously allowing for various interpretations across the multiple areas of the testing process in a specific setting. This conceptualization of test fairness helps to clarify the scope of fairness investigations, emphasize the importance of stakeholders’ involvement, and recognize the role of value judgments that are embedded in the testing situation.

A variety of specific aspects have been identified for fairness investigations in Willingham’s work (Willingham & Cole, 1997; Willingham, 1999). Since those specific aspects are mainly written for educational researchers and the technical community, other stakeholders such as test users and test takers are not involved in all of the aspects as they do not always have experience or knowledge. Hence adaptation is needed in using his framework in this study. To guide my fairness investigations in action, I developed an operational model that focuses on six aspects of Willingham’s framework. These six aspects are consistent with the research context and the purpose of the study. The following model (see Figure 2) provides a visual representation of Willingham’s conceptual framework.
The operational model, from the inner to the outer circle, includes: comparability for groups and individuals of test takers (comparability), six specific aspects (see boxes) within the four areas of the testing process (process), three groups of test stakeholders (participants), and the GSEEE with arrows indicating the dynamics of stakeholders’ value judgments in the testing context (situation). The six specific aspects examine comparable treatment for both group and/or individual test takers. Guided by the fairness literature, some of the six aspects focus on comparability investigations for test taker groups (1.1 test bias and 4.1 predictive bias) and some examine comparability for individuals (2.1 standardized administration, 2.2 test accommodations, and 3.1 scoring.
procedure). Based on the literature, the aspect of decision-making (4.2) is examined for both groups and individuals of test takers. Considering the demographic information of China (Postiglione, 2006), I am particularly interested in two grouping variables: gender (Male vs. Female), academic background (Humanities and Social Sciences vs. Sciences).

The groups of test users, developers, and takers are included in the operational model. Although test developers’ perceptions are not examined in this study, the test developer group is included in the model due to their role. Also the three groups of stakeholders do not necessarily correspond only to the areas that they fall within. For example, the test user group often provides meaningful information in terms of the fairness of score-based test use, as shown in Figure 2. Groups other than test users, for example, test takers, may also address issues of fairness in test use. In the following, I elaborate on the six specific aspects across the four areas of the testing process.

Area 1: Design and development

1.1 Test bias: this form of investigation examines if the GSEEE test taker groups (e.g., gender and academic background) obtain comparable opportunities in terms of test construct, format, and content on a given test administration, so that they can demonstrate their knowledge and skills that are relevant to the purpose of the GSEEE.

Area 2: Administration

2.1 Standardized administration: this form of investigation examines if individuals GSEEE test takers without disabilities take the GSEEE under uniform physical conditions and consistent and secure administration.
2.2 Test accommodations: this form of investigation examines if GSEEE test takers with disabilities and special needs are provided with accommodations so that their relevant knowledge and skills can be accurately demonstrated.

Area 3: Scoring

3.1 Scoring procedure: This form of investigation examines if the performance of individual GSEEE test takers is accurately and consistently scored to demonstrate relevant knowledge and skills that they have acquired.

Area 4: Score-based use

4.1 Predictive bias: This type of investigation examines if the GSEEE is an effective and accurate predictor of subsequent learning for the GSEEE test taker groups (e.g., gender and academic background).

4.2 Decision-making: This type of investigation examines if decisions with the GSEEE test takers, both groups and individuals, based on the GSEEE test scores along with other relevant information are made appropriately.

Stakeholders’ perceptions of the fairness of the GSEEE are examined based on their knowledge and experience about policies, guidelines, and practices adopted in the GSEEE. In the GSEEE, what test takers know and can do is demonstrated through test items (e.g., multiple-choice items in reading comprehension on a topic related to biology) in a given administration. After that, test takers’ performances are scored (e.g., automated scoring) and used for a specific purpose (e.g., admission purpose). Unfairness may be perceived and introduced at any of these areas, regardless of the care that may have been exercised in other areas. Empirical results based on stakeholders’ perceptions may help to shed light on fairness concerns. Stakeholders’ general conclusions of the
fairness of the GSEEE may also help to identify key elements in the conceptualization of test fairness. Although test users’ and takers’ concerns about fairness may sometimes be exaggerated or inconsistent with psychometric results, their perceptions provide an alternative lens and help to identify problems related to the test.

Empirical Studies on Stakeholders’ Perceptions of Test Fairness

Empirical research on test fairness based on stakeholders’ perceptions has long been conducted in some fields; for example, job applicants’ test-taking experience (Bacdayan & Geddes, 2009; Truxillo, Bauer, & Sanchez, 2001). More than a decade ago, Gilliland (1993) put forward his influential model of the perceived fairness of selection systems. Grounded on his model, Bauer et al. (2001) developed a 5-point Likert Selection Procedure Justice Scale (SPJS) using exploratory and confirmatory factor analyses, and 11 factors were extracted. Some of these factors are similar to the six aspects outlined in my operational model, for example, standardized administration (see Figure 2). The SPJS has been widely used in personnel management, business, and medical education (Patterson et al., 2011; Pepper & Pathak, 2008). The theoretical, empirical, and methodological development on perceived fairness in these disciplines provides evidence about the usefulness of the stakeholder approach in fairness research.

Although the stakeholder approach has been used in language testing research, empirical studies are often undertaken within test validity frameworks, not fairness frameworks (Cheng & DeLuca, 2011; Karelitz, 2012). Fairness and validity are closely related (Willingham & Cole, 1997); however, they involve “different emphases” (Kane, 2010, p. 78). Since the 20th century, test validity theories have been well developed (Thorndike, 1916; Cronbach, 1988; Kane, 2006; Messick, 1989). Subsequently, empirical
studies are often conducted within test validity frameworks, and fairness has typically been addressed as a secondary theme within validity theories. Empirical studies on stakeholders’ perceptions within fairness frameworks are limited. In the following, I review studies that provided information related to test fairness from stakeholders’ perceptions using fairness or validity frameworks in language testing. These studies discussed perceptions (Karelitz, 2012), reflections (Cheng & DeLuca, 2011), experience (Fox & Cheng, 2007), attitudes (Coleman, Starfield, & Hagan, 2003) and response to the test (Brown, 1993). They were conducted with a wide range of test stakeholders including test takers (Tsai & Tsou, 2009), teachers (Du, 2007), administrators (Hyatt, 2012), raters (May, 2009), educators (Schulte, Elliott, & Kratochwill, 2000), and parents (Scott, 2007). The majority of these studies used qualitative methods such as interviews, open response questionnaires, and observation.

Empirical studies have examined perceptions across various aspects of the testing process from test takers’ and teachers’ perspectives (Cheng & DeLuca, 2011; Du, 2006; Karelitz, 2012). Drawing on Messick’s validity framework (1989), Fox and Cheng (2007) compared interpretations of two groups of test takers — English as a first language (L1) and as a second language (L2) regarding test constructs and interactions among test design, interpretations, and accounts of classroom practices in a Canadian literacy test. The study conducted 33 focus group interviews that included 22 L1 and 136 L2 students. Perceptions on fairness were evident from the test takers’ comments, particularly in test accommodation, preparation, and administration. Interestingly, L2 focus group participants stated that it was unfair to be treated differently through accommodation
practices because such practices were not offered in classroom assessment. In the study, focus group interviewing proved to be a valuable tool to elicit stakeholders’ perceptions.

While the above studies examined perceptions across multiple aspects of the testing process, most research focused on one specific aspect of the testing process. These studies investigated test takers’ and users’ perceptions in terms of, for example, potential test bias (Fox, 2003), standardized administration procedures (Huang & Garner, 2009), appropriateness and fairness of test accommodations (Schulte, Elliott, & Kratochwill, 2000), scoring practices (May, 2009), and fairness and ethicality of test score uses (O’Loughlin, 2011). Among these studies, Fox’s study (2003) is insightful because it used the fairness literature to guide investigations and detected test bias based on stakeholders’ perceptions. Her study adopted an “ecological approach through the analysis of test taker and rater responses” with intent to “identify bias that systematically undermines the measurement of the ability of a ‘group’ of test takers” (p. 26). Using open response questionnaires, the study compared the accounts of test takers (n=423) and raters (n=12) on two versions (anchor vs. new) of a high-stakes language test. Results found that the new version created fairness concerns regarding the topic. However, the test takers who expressed satisfaction or dissatisfaction with the test topics were not of a particular gender, linguistic, cultural, academic background, or language proficiency group. Ultimately, the perceptions of the test takers and raters resulted in the redefinition of the test specifications. As Fox (2003) pointed out, the study identified fairness concerns that might otherwise remain undetected by traditional methods such as DIF. The study also pointed out the challenge in identifying test taker group in bias research. From
the perspective of these test takers and raters, unfairness may be related to any systematic error, but not necessarily for test taker groups with identifiable traits.

Overall, the research on stakeholders’ perceptions of test fairness pinpoints four issues. First, through the accounts of test stakeholders in a particular context, it is possible to identify issues and concerns that may be undetected by traditional psychometric approaches. Test users’ and takers’ involvement is helpful, particularly in a situation where limited information regarding the fairness of the test is available to the public. Qualitative methods such as interviewing generate rich, emic information about test fairness. Second, since not all stakeholders are involved in all the aspects of the testing process, it is important to identify appropriate stakeholders in fairness investigations in relation to specific aspects of the investigations. While some stakeholders, such as test takers, may address fairness issues across various aspects of the testing process, others, such as administrators and parents, may provide meaningful information only on some aspects, for example, test use. Third, unfairness can be perceived in any aspect of the testing process. Stakeholders’ perceptions can contribute to identifying possible concerns and improving tests and test specifications. Fourth, the results and recommendations as suggested by the studies reviewed here are closely related to practices used in specific contexts.

As Crocker (2003) suggested, investigations of stakeholders’ perceptions expand our understanding of fairness in useful ways. Systematically valuing and incorporating stakeholders’ perceptions allows us to make more convincing arguments about the accuracy and appropriateness of the inferences we draw from tests, ultimately leading to increases in test fairness. Given the social, economic, and educational changes of recent
years in China, examining stakeholders’ perceptions of test fairness is important, and questions related to the fairness of the GSEEE should be examined empirically.

Method

In this qualitative study, I used a multiple perspective approach to examine the fairness of the GSEEE. Prior to data collection, the research received clearance from the Queen’s University General Research Ethics Board and complied with the Tri-council Standards on Research with Human Subjects (see Appendix B for the Clearance Letter). In addition, all the participants were given a letter of information and signed a consent form before interviewing (see Appendix G, H, I, J, K, & L).

Participants

The study was conducted with three groups: graduate program administrators, and teachers who taught English at master’s programs, and test takers. Five graduate program administrators and five English teachers were recruited from four universities through personal connections. The five graduate program administrators were chosen based on the criteria that they were the key officers of graduate programs and they were involved in making crucial decisions regarding entrance to master’s programs based on the applicants’ GSEEE test scores and other measures. They were either the dean or associate dean of a graduate school/graduate program. It was found later during the interviews that the participants also worked as master’s advisors and were actively involved in teaching master’s students. The administrators, four male and one female, ranged in age from their 40s to 60s. All the administrators had been involved in graduate schools/programs for more than 3 years.
Five teachers were selected on the conditions that they taught English courses at the master’s level and played a leading role in classroom teaching and learning in their units/schools. They served as unit head or dean of the School of Foreign Languages. During interviewing, it was found that four of them were also involved in GSEEE rating and one had additional experience in GSEEE item writing. Although these teachers served as raters or item writers, they considered English teaching as their primary responsibility. The English teachers, three male and two female, ranged from their early 30s to early 50s. They had various lengths of teaching experience at the master’s level, with the shortest being three years and the longest more than 20 years.

Test takers were recruited from one large university in North China through posters in the public area on campus (Appendix M). This university, with over two hundred master’s programs, has attracted a large number of applicants across China each year. Those who took the 2012 administration of the GSEEE were invited to participate in focus group interviews. Five focus group interviews, involving 20 test takers in total, were conducted, with each group composed of 2 to 6 test takers. Table 8 provides specific information related to the interviewed test takers. To be consistent with Willingham’s work and the operational model, the focus group participants were recruited using a purposeful sampling technique. Among the 20 test takers, there were 10 males and 10 females, who came from a variety of academic backgrounds. The test takers were representative of the groups of gender and academic background. However, the recruitment of test takers with disabilities was not successful; not a single disabled test taker was identified. The 20 test takers reported a wide range of overall and sub-skills (reading, writing, speaking, and listening) English proficiency levels from 1 (the lowest)
to 3 (the highest) based on their own judgment. They ranged from 23 to 28 years old, and their GSEEE test scores varied from the highest 77 to lowest around 30. Among the 20 participants, 8 students were accepted into master’s programs. A majority of the test takers, 17 in total, took the GSEEE only in 2012, while the other three sat the GSEEE more than once in addition to the 2012 administration (e.g., the 2011 administration).

Table 8

*Background Information and Admission Status of the Test Takers*

<table>
<thead>
<tr>
<th>Focus group</th>
<th>Test takers</th>
<th>Gender</th>
<th>Age</th>
<th>Academic background</th>
<th>Test score</th>
<th>Pass/fail</th>
<th>Admission status</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>S1</td>
<td>M</td>
<td>24</td>
<td>Computer Science</td>
<td>NA</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>M</td>
<td>24</td>
<td>Computer Science</td>
<td>51</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>F</td>
<td>25</td>
<td>Economics</td>
<td>59</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>F</td>
<td>24</td>
<td>Computer Science</td>
<td>NA</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td>S5</td>
<td>M</td>
<td>28</td>
<td>Chinese Pharmacy</td>
<td>39</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S6</td>
<td>F</td>
<td>24</td>
<td>Political Science</td>
<td>59</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td>G3</td>
<td>S7</td>
<td>F</td>
<td>23</td>
<td>Economics</td>
<td>49</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S8</td>
<td>M</td>
<td>24</td>
<td>Computer Science</td>
<td>50+</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td></td>
<td>S9</td>
<td>F</td>
<td>24</td>
<td>Mathematics</td>
<td>64</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S10</td>
<td>M</td>
<td>23</td>
<td>Mathematics</td>
<td>61</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td>G4</td>
<td>S11</td>
<td>F</td>
<td>24</td>
<td>Chinese Literature</td>
<td>74</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S12</td>
<td>F</td>
<td>23</td>
<td>Communication Engineering</td>
<td>60+</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td></td>
<td>S13</td>
<td>F</td>
<td>24</td>
<td>Chinese Literature</td>
<td>30+</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S14</td>
<td>M</td>
<td>24</td>
<td>Civil engineering</td>
<td>NA</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S15</td>
<td>M</td>
<td>23</td>
<td>Communication Engineering</td>
<td>40+</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S16</td>
<td>M</td>
<td>23</td>
<td>Civil engineering</td>
<td>45</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>G5</td>
<td>S17</td>
<td>M</td>
<td>26</td>
<td>Electric Engineering</td>
<td>67</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td></td>
<td>S18</td>
<td>F</td>
<td>23</td>
<td>Communication Engineering</td>
<td>77</td>
<td>P</td>
<td>Accepted</td>
</tr>
<tr>
<td></td>
<td>S19</td>
<td>M</td>
<td>25</td>
<td>Electric Engineering</td>
<td>NA</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S20</td>
<td>F</td>
<td>23</td>
<td>E-Business</td>
<td>54</td>
<td>P</td>
<td>Accepted</td>
</tr>
</tbody>
</table>
Instruments

The study used one-on-one interviewing with administrators and teachers and focus group interviewing with test takers. One-on-one interviews and focus group interviews were semi-structured in nature. The interview protocols were used to probe for clarification, solicit greater detail, and explore participants’ perceptions in depth. Essentially, the interview protocol was designed based on Willingham’s conceptual framework and the operational model. Given the participants’ limited English proficiency, all of them were interviewed in their first language — Mandarin. Because English is the language used in Willingham’s framework and the operational model, the interview protocols were first developed in English, and then translated into Chinese. The translation was verified by a bilingual researcher.

Based on the literature, not all stakeholders were involved in all the aspects of the testing process. While test takers may address fairness issues across various aspects of the testing process (Cheng & DeLuca, 2011), others, such as administrators, may provide meaningful information only in some aspects, for example, test use (O’Loughlin, 2011). As such, the interview protocol for test takers included all the six aspects of the testing process as outlined in the operational model (see the previous section), and the protocols for administrators and English teachers mainly focused on test use. However, the interview protocols provided administrators and teachers with opportunities to express their perceptions on the other aspects of the testing process (e.g., scoring) in which they expressed they had knowledge and experience, and were able to provide rich information.

Specifically, the interview protocols for graduate program administrators and English teachers also included three sections (see Appendices N and O): their background
information and work/teaching experience (Q1-Q3), experience and knowledge of the GSEEE (Q4 & Q5), and perceptions of the fairness of the testing process (e.g., score-based use) and overall perceptions (Q6 - Q9). The interview protocol for test takers included three sections (see Appendix P): first, their background information and their perceived English proficiency level (Q1-Q3); second, their learning experience and knowledge of the GSEEE (Q4 & Q5); and third, their perceptions of the fairness of all the six aspects across the 2012 testing process as well as their overall perceptions of the fairness of the GSEEE (Q6-Q11).

Data Collection and Analyses

The one-on-one interviews with graduate program administrators and English teachers were conducted in February and March of 2012. The interviews ranged from 30-60 minutes and were voice-recorded. While the questions focused on the respondents’ perceptions of score-based use, the interviews provided the administrators and teachers opportunities to express their perceptions in the other aspects of the testing process based on their experience and knowledge.

The test taker interviews were conducted in two time periods. First, focus group interviews were conducted within a month after the test takers completed the GSEEE in January and February of 2012. Each focus group interview ranged from 60-90 minutes and was also audio recorded. Focus group interviews covered all the interview questions except scoring (Q8) and score-based use (Q9). Then, individual follow-up phone interviews, about 3-5 minutes long, were completed in May 2012 after the test takers had been informed of their GSEEE scores and admission status. Interview questions focused on scoring (Q8) and test use (Q9).
Data analysis consisted of a multi-step, iterative process to minimize bias (McMillan and Schumacher, 2006). A Research Journal was maintained to enhance the credibility of the findings and establish an audit trail through memo writing (McMillan and Schumacher, 2006). Nvivo 8 was used to analyze the interview data, and three sources of data were analyzed together. To organize and prepare the data for analysis, I transcribed them verbatim in Chinese. Subsequently, I translated them into English and the translation was verified by a native speaker with the academic background. After that, I used a combined technique of deductive and inductive thematic analyses to interpret the interview data (Fereday & Muir-Cochrane, 2006). The process incorporated the deductive a priori template of themes approach outlined by Crabtree and Miller (1999) and the data-driven inductive approach of Boyatzis (1998). The initial list of themes was based on Willingham’s framework and the operational model. I then modified them within the process of the analysis as new themes emerged inductively. Through this hybrid process of data analysis, I could not only identify themes that were integral to the process of deductive thematic analysis, but also discover themes that emerged directly from the data using inductive coding.

Results

The following presents an overview of five themes identified/emerged from the interview data: item quality, standardized administration, scoring, cut-scores and admission decisions, and overall perceptions of the GSEEE. Each theme was broken down into a number of codes to illustrate the participants’ perceptions of the fairness of the GSEEE. Table 9 shows the code list and frequency counts per code with each of the source groups. While some themes and codes were prominent and evident across all the
three groups, some were eminent only within one group. In addition, divergences occurred. In the following, I elaborate on each of the themes.

Table 9

*Themes and Code Frequencies*

<table>
<thead>
<tr>
<th>Themes and codes</th>
<th>Source and frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GA</td>
</tr>
<tr>
<td>Theme 1: Item quality</td>
<td>68</td>
</tr>
<tr>
<td>Code 1: Constructs being tested</td>
<td>26</td>
</tr>
<tr>
<td>Code 2: Item difficulty and discrimination</td>
<td>52</td>
</tr>
<tr>
<td>Theme 2: Standardized administration</td>
<td>26</td>
</tr>
<tr>
<td>Code 1: Measures to ensure standardization</td>
<td>13</td>
</tr>
<tr>
<td>Code 2: Systematic and random irregularities</td>
<td>17</td>
</tr>
<tr>
<td>Theme 3: Scoring</td>
<td>25</td>
</tr>
<tr>
<td>Code 1: Rater effects on scoring</td>
<td>12</td>
</tr>
<tr>
<td>Code 2: The writing template</td>
<td>17</td>
</tr>
<tr>
<td>Theme 4: Cut-scores and admission decisions</td>
<td>61</td>
</tr>
<tr>
<td>Code 1: Equal treatment</td>
<td>29</td>
</tr>
<tr>
<td>Code 2: Differential treatment</td>
<td>11</td>
</tr>
<tr>
<td>Code 3: Multiple measures for the final decision-making</td>
<td>23</td>
</tr>
<tr>
<td>Code 4: The use for admission and other uses</td>
<td>8</td>
</tr>
<tr>
<td>Theme 5: Overall perceptions of the GSEEE</td>
<td>26</td>
</tr>
<tr>
<td>Code 1: The fair testing process</td>
<td>14</td>
</tr>
<tr>
<td>Code 2: The testing-centred tradition</td>
<td>9</td>
</tr>
<tr>
<td>Code 3: Usefulness</td>
<td>7</td>
</tr>
<tr>
<td>Code 4: The merit-focused belief</td>
<td>12</td>
</tr>
<tr>
<td>Code 5: Practicality</td>
<td>5</td>
</tr>
</tbody>
</table>

*Note:* GA—the administrator group, and ET—the English teacher group, and TT—the test taker group. Theme frequencies do not equal the total of code frequencies due to double-coding of data.
Item Quality

The theme of item quality was identified only among the test takers due to their test-testing experience. The administrators and teachers expressed their limited knowledge, experience, and interests, as an administrator commented “we have never seen or read the GSEEE because it is only the test takers who are going to take the test” and a teacher stated “the GSEEE scores were separate from English teaching in the classroom.” In contrast, the test takers had extensive discussions on test items and their quality. They provided detailed information concerning the GSEEE, such as timing, length, constructs, format, and the weight of each item/section. They obtained such information from past examination papers, test preparation materials, and coaching programs. Interestingly, while the interview questions were intended to elicit information regarding potential item bias for the GSEEE test taker groups (gender or academic background), the test takers expressed their thoughts about the test for individuals rather than groups. Those who expressed satisfaction or concerns were not of a particular gender or academic background. Their perceptions focused on two areas: (1) constructs being tested and (2) item difficulty and discrimination.

The choice of constructs being tested emerged as one aspect; however, divergences existed among the test takers. As mentioned earlier, the GSEEE only tests reading and writing; listening and speaking are tested once test takers enter the second round of re-examinations. Some test takers endorsed this testing practice, saying that it was fair to include only reading and writing tasks in the GSEEE. They believed that the inclusion of listening and speaking tasks would require “much greater financial support” and that China was still constrained by “limited resources.” In contrast, other test takers
felt that the potential unfairness arose due to construct under-representation. They believed that the GSEEE was “a comprehensive test” and the inclusion of listening and speaking items might increase their chances to “enter the second round of re-examinations.” As Willingham and Cole (1999) stated, narrowly defined constructs usually measure some aspects of the construct, but not others; as a result, it is likely to advantage some test takers and disadvantage others. However, the perceptions on the constructs being tested in the GSEEE were mixed because the test takers had different foci. While some test takers believed that fairness could be compromised due to practicality, some expressed their fairness concern because of lack of construct representativeness in the GSEEE.

The test takers also complained of the high difficulty level of the GSEEE, and concluded that the GSEEE did not reflect differences of test takers’ proficiency levels. The GSEEE, a norm-referenced test with a total score of 100, is designed for non-English majors in all disciplines of Humanities, Social Sciences, and Sciences (He, 2010). The participants were suspicious of whether the GSEEE could distinguish test takers with different proficiency levels. The following test taker stated his view of how the high difficulty level might impact differentiation levels and influence fairness:

I feel that the difficulty level is very high with reading texts, hence, there is much guessing. To give an example, there are two test takers, one having a proficiency level of 50% and the other 40%. If the text in the GSEEE is designed for those who must score at least 80% to prove comprehension, chances are that there are not major differences between the GSEEE scores of the two test takers. In other words, if texts can be designed with a more appropriate difficulty level, regardless
of test content or vocabulary complexity, students at different levels from 10% to 80% would be more accurately distinguished. Currently, the difficulty level is way too high. It is hard to distinguish the real differences among those at medium and low proficiency levels.

The test takers believed that the GSEEE test developers intended to “trick” students through negative wording or emphasizing unimportant details. The text for translation normally consisted of long sentences that were syntactically and semantically complex. As a result, the test takers reported that there were “many test-taking strategies” and “much guess.”

*Standardized Administration*

Again, only the test takers elaborated on their perceptions of the fairness of the GSEEE’s administration because of their test-taking experience. Responses delineated two aspects in test administration: measures to ensure the standardization of administration as well as systematic and random irregularities. None of the test takers discussed any meaningful information related to test accommodations

The test takers reported that many policies, guidelines, and practices were used to ensure standardization, warrant test security, and combat cheating. The NEEA used various technologies to ensure test standardizations, such as “the signal-free zone,” “the security check at the entrance,” “the signal detection device at the door of each classroom,” and “the random re-checkup of test takers’ ID.” The test takers were aware of the significant consequences of violating testing regulations such as “receiving a zero score,” “being

---

14 One of the teachers, who was also an item writer, explained that, to her understanding, there were no accommodations provided for test takers with special needs, not even *Braille* for the visually impaired. The information was confirmed by an official from a provincial NEEA branch.
ineligible to attend the test in the following year,” and “being reported to one’s undergraduate program for penalties.”

Regardless, the test takers believed that irregularities and dishonesty were visible, as in the Chinese saying: “while the priest climbs a post, the devil climbs ten,” meaning that cheaters can always find ways to bypass testing regulations regardless of how much effort has been made to ensure security. The test takers reported that “unusual or out-of-the ordinary circumstances happened all the time.” They described occurrences of both systematic and individual irregularities, with random issues cited more frequently. The test takers reported that the essay topic of the 2012 administration was disclosed in advance online in some provinces, which involved a large number of test takers; this occurrence was later confirmed by the government officials (INFZM, 2012). Some test takers complained about the tight proximity of seats and distractive testing conditions; one test taker mentioned that the proctor in her testing room fell asleep; and another described his success in peeking at another’s test answers.

**Scoring**

Two aspects in terms of scoring were identified among the test takers and teachers: rater effects on scoring and the writing template. Due to their rating experience, four among five teachers described their knowledge and expressed their perceptions on the fairness of the rating practices. First, while some scoring policies and practices were considered unfair among the test takers and teachers, some were praised as evidence of fairness. The test takers reported their anecdotal observations about scoring inconsistencies between one province and another. The observations were echoed by the teachers who had opportunities to be involved in the GSEEE scoring. These teachers
talked about their concerns regarding different interpretations of the essay scoring scheme and potential problem using one single rater in rating essays, which led to inaccuracy about test takers’ ability.

The test takers and teachers also discussed scoring policies, guidelines, and practices that they believed to enhance fairness. One testing practice received much endorsement: all test scripts from test takers who applied for the same university, no matter where they took the test in China, were sent to the same rating center for scoring. Further, two teachers commented on a recent new practice—rating using the computer, which they believed made the scoring “fairer” and reduced “the percentage of papers inaccurately rated.” Machine scoring for dichotomously scored items were used for many years in the GSEEE; however, rating constructed-response items such as the essay on the computer just started in some provinces. The teachers introduced that all essays were first scanned and put into the computer, and then assessed by raters. The teachers believed there were several benefits. First, rating on the computer provided clear information regarding inter-rater and intra-rater reliability, so issues related to individual raters could be identified instantly. Second, it provided quality control for every single writing task. In comparison with the previous practice where only a small portion of randomly selected test papers were reviewed, rating on the computer reduced the likelihood of an essay being mistakenly scored or reported. Third, compared with the previous one reading of writing tasks, rating on the computer required two readings and the final score would be the average of the two scores. If a large score discrepancy existed within the scores on a test paper, a third reviewer would be invited to make an additional judgment. If still unsolved, the paper would be submitted to a leading supervisor for a final judgment.
Rating using the computer was considered to “significantly” improve scoring quality and enhance test fairness.

The second aspect emerging from the interview data in scoring was the use of writing templates by test takers in the GSEEE essay writing. A writing template refers to the memorization of key words and topic sentences to be used to provide a vague discussion of a social problem or phenomenon (He, 2010). The use of definite articles, adjectives, determiners, and pronouns seems to be relevant to a particular issue, but in fact it is like an all-purpose adhesive that can be applied to a variety of similar topics. The test takers reported widespread use of the writing template, and the teachers expressed their difficulty in scoring these essays. The test takers reported that they memorized “around a dozen templates” before taking the GSEEE, similar to the preparation for the “eight-legged essay \(^{15}\)” in the China Imperial Examinations; during the GSEEE, the test takers chose which one was closest to the topic being tested, similar to “filling in the blanks,” in combination with a few words of their own; and they felt “good” about themselves after the test. The writing template was reportedly easy to obtain through various sources including formal printed testing preparation materials, coaching programs, and online testing preparation websites for the GSEEE.

Although widely used among the test takers, no consistent scoring practices were reported regarding how to identify writings using templates as well as how to score those writings. Different attitudes and practices emerged among the teachers in how essays using templates were scored. Some teachers gave “an extremely low score” and some gave “a minimal score considering test takers’ efforts to memorize the sentences.” One

---

\(^{15}\) The eight-legged essay (baguwen) is a written form of argument in the Imperial Examination in China. This format later becomes synonyms of pedantry or triteness.
teacher/rater explained the practice in the provincial NEEA branch where she worked:
“we want our raters to focus on the part which test takers wrote themselves and rate the
writing based on those sentences. [They should] Try to diminish the effectiveness of the
template.”

_Cut-scores and Admission Decisions_

While predictive bias did not emerge as a theme, cut-scores and admission
decisions is a theme that emerged strongly from all the three data sources. Four aspects
were identified from the interview data: (1) equal treatment, (2) differential treatment, (3)
multiple measures in admission decisions, and (4) the use for admission and other uses.
First, the administrators, teachers, and test takers endorsed the testing policy that the
MOE decides the national GSEEE cut-scores, which is used for a majority of institutes,
over 900 of them, across the whole country. The administrators, in particular, elaborated
on this policy and explained that cut-scores are related to the quota system—Enrollment
Planning Objectives. The participants believed that the national cut-scores helped to
achieve fairness: “everyone is equal in front of the test scores.” Since the national cut-
scores established a minimum requirement, “those who are very poor” could not enter the
second round. They felt that minimum cut-scores and the ranking system helped
universities to fight against “public bribery”, “corruption”, and “fraud.”

Second, less frequently (see code frequencies in Table 9), the administrators,
teachers, and test takers also discussed two types of differential treatments: independently
determined cut-scores among the 34 top universities, and cut-score differences based on
ethnicity and region. The participants, administrators in particular, endorsed the policy of
differential treatment among 34 universities, which they believed empowered institutes in
decision-making. The administrators accepted this practice as “the best practice of the profession” because it expands the universities’ admission autonomy, despite that these programs still had “limited choices” due to the use of the quota system. The administrators asked for more freedom so the universities could respond to the needs of the local educational system.

The administrators, teachers, and test takers also discussed the fairness of cut-score differences based on ethnicity and region. According to the testing policies and practices, ethnic minorities and those who apply for programs in less developed regions (mostly in minority-populated areas) are given preferential treatment with lowered cut-scores (Wang, 2009). The stakeholders considered this policy to be fair since there existed “differences in educational opportunities.” The participants commented that differential treatments were implemented considering the gaps of ethnic differences and east-west economic disparities in terms of learning opportunities and educational development. They took into account the interests of those select groups.

Third, the three groups of the participants, the administrators in particular, described that the use of various measures including the GSEEE and other criteria for final decision-making was “comprehensive and fair.” The administrators described the importance and use of various measures in deciding which test takers can be admitted into master's programs. In addition to the GSEEE and the other three scores in the GSEE in the first round, the measures in the second round included: (1) a written comprehensive exam, (2) a listening and/or writing exam organized by the School of English, (3) an interview on oral English and subject knowledge by the re-examination committee, (4) the scores in College English Test 4/6, (5) the undergraduate’s GPA, and (6) a physical
examination. Test takers were evaluated, as one administrator described, based on “three or four” types of English scores (including the GSEEE) to demonstrate their English proficiency. “A total weighted score” based on the previous indexes was used to decide who could enter master’s programs. Universities and their programs independently determined the weighting of various indexes by considering the training requirements of their master’s students. The decision-making process using the GSEEE and other measures was considered as “a holistic, objective, and fair practice.”

Fourth, the administrators, teachers, and test takers discussed the use of the GSEEE for the admission purpose and other purposes. The participants concentrated on the “selection” purpose and emphasized the importance of “the gatekeeping role.” They attached much more importance to the societal use of the GSEEE to stream and weed out test takers than the purpose of evaluating English proficiency. The teachers also talked about the other uses of the GSEEE test scores such as placement, whether students can waive English credits partially, or go to combined “Master+PhD” programs. The reasons to use the GSEEE for these purposes included “time constraints in placing students in the first week of the academic year” and “other universities have used [this practice] for many years.” The other uses of the GSEEE scores seem to be arbitrary and lack empirical evidence, as one teacher pinpointed the problem that his institute encountered – “The cut-score based on the GSEEE, 70%, to decide whether certain credits can be waived is the same this year as the last year; however, many more students passed the cut-score this year than last year. So we have to raise the score to 75%.”
Overall Perceptions of the Fairness of the GSEEE

Despite concerns about the fairness of some policies, guidelines, and practices, the administrators, teachers, and test takers believed that, overall, the GSEEE was a fair test. The GSEEE should continue to be used, progress could be gradually pursued, and fairness would be improved further in the long run. There were several reasons for their overall conclusion; these are presented below in descending order based on the total coding frequencies (see Table 9): (1) the fair testing process, (2) the testing-centred tradition, (3) usefulness, (4) the merit-focused belief, and (5) practicality. First, the administrators, teachers, and test takers believed that the GSEEE was fair because of its policies, guidelines, and practices that intended to provide test takers with equal opportunities to take the same test at the same time using the same administration procedure, under the same scoring scheme. And admission decisions were made primarily based on score ranking. The participants, the administrators in particular, also added that it was important to use different cut-scores to compensate those who were in a disadvantaged position such as ethnic minorities. In general, what they believed to be a fair test remained consistent with what the testing policies and practices promoted. The participants claimed there was “no better alternative choice” to replace the GSEEE in current China.

Second, the administrators, teachers, and test takers reported that they accepted the GSEEE due to the influence of the testing-oriented tradition. They pointed out that high-stakes testing had been used for thousands of years in China ever since the Imperial Examinations, and the heavy weight of testing in the overall education system was evident. They stated that “people have become accustomed to it.” The participants
described the existence of the GSEEE as “just like the existence of an apple.” They believed “everything in existence was reasonable” and that the test was immune to critique and reflection. They expressed their sense of “affective commitment” and showed acceptance of the GSEEE’s configuration.

Third, the GSEEE was perceived to be fair also because it achieved the primary goal of the GSEEE—the screening purpose. As one administrator described, there were “more than 6000 applicants applying for master’s programs every year at my university”. The GSEEE provided a quick solution to screen test takers and decide who might enter the second round of re-examinations. The societal use of the GSEEE to stream and weed out test takers was highly valued while the purpose of English proficiency was often ignored among the administrators, teachers, and test takers. The GSEEE was seen by the stakeholders as a useful, efficient way to achieve the major goal of selection.

Fourth, the administrators and teachers highlighted their merit-based belief, which influenced their overall perception of the GSEEE as a fair test. They believed that the GSEEE, first and foremost, was to provide a level playing field for all test takers to demonstrate their merits and worth. Those who received high test scores would show their merits, and subsequently, should be selected for higher education opportunities. The GSEEE promoted fairness based on “performance” instead of individual needs, wealth, family connections, or political affiliation. The GSEEE provided such a chance of “upward mobility” for anyone, especially those who came from low socio-economic background. As one administrator explained:

The only thing you can fight is this cut-score. And once the cut-score is drawn, you are either in or out. No one can help you. For those who do not have any
background, this is the only chance they can climb up. These kids, due to their poor family background, may not have a chance to develop strong, comprehensive ability; for example, they may not be versatile and do not receive training in piano, painting, etc. However, this test gives them an opportunity. And they may climb up and be successful in the future if given this chance. After all, there is no better choice since there is a huge gap between the rich and poor in China now.

Fifth, the administrators, teachers, and test takers took practicality into account as one reason for their conclusion that the GSEEE was a fair test. They explained that China is a large country and the GSEEE is administered to “more than one million test takers.” They argued that the existence of the GSEEE was reasonable due to the “cost-benefit” consideration. Despite “not being absolutely fair,” the participants commented that the MOE “took one step after another” toward the path of fairness, such as the Region A & B classification and the autonomy of 34 top universities to decide their own cut-scores. These changes were believed to be consistent with the educational, economic, and social development of the current situation in China.

Discussion

This study explored administrators’, English teachers’, and test takers’ perceptions of the fairness of the GSEEE. Although these reports represent the perceptions of only five graduate program administrators, five English teachers, and 20 test takers, the major themes identified/emerged from the interviews reflect the complexity and intricacy of the fairness of the GSEEE. In the following, I present a focused discussion to answer the three research questions: stakeholders’ knowledge of
the GSEEE, their perceptions of the fairness of the GSEEE policies, guidelines, and practices, and their overall perceptions of the fairness of the GSEEE.

Knowledge of the GSEEE

The interview results found that the program administrators, English teachers, and test takers had different levels of knowledge about the GSEEE. While the test takers provided much information related to all the six aspects of the GSEEE process, the administrators and teachers provided information only in select areas. Such results are not surprising because the three groups had different experience and interests. Even though the test specifications are updated and published by the NEEA every year and contain important information such as test purpose, constructs, timing, length, weight of each section, formats, scoring, and a vocabulary list, test users, administrators in particular, lacked information about the test itself. Although the standards, guidelines, and codes of practice generally recommend that test users should review and evaluate the fundamental information regarding the test and its quality (ILTA, 1999), whether test users (e.g., administrators and teachers) actually follow these recommended practices is a different issue. Similar results have been found in other high-stakes language testing. In the study conducted by Coleman, Starfield, and Hagan (2003), it was found that students were more knowledgeable than staff on a wide range of themes related to the IELTS. In Hyatt’s study (2012), the vast majority of administrators admitted that “they did not really have a clear understanding of the content and process of the IELTS test” (p. 12). Academic admissions officers had limited knowledge about the IELTS and the meaning of the test scores.

16 The GSEEE test specifications include a list of approximately 5,500 required words and idioms.
In the GSEEE context, the graduate program administrators and teachers did seem to be interested in test items and quality, partly because the GSEEE is mandatory and test users are not in a position to select another test, as one respondent expressed as “unable to make a choice”. Because the GSEEE primarily plays a societal role by weeding out test takers and the test is separate from English teaching in the classroom, the teachers showed limited interest in obtaining relevant information about the GSEEE. In the absence of a sound knowledge, the GSEEE has been adopted for unwarranted uses such as placement. Another possible reason for lack of interest could be that test users such as administrators and teachers are unaware of the importance of item quality and have rarely been provided with review materials or technical reports, as was pointed out to the researcher by one of the officials working at the provincial NEEA branch.  

Perceptions of the Fairness of the GSEEE Policies and Practices

The results showed mixed perceptions of the fairness of the GSEEE testing policies, guidelines, and practices from the perspectives of administrators, teachers, and test takers. While some policies and practices were perceived as being fair (e.g., measures to ensure standardization, equal treatment), some provoked concerns (e.g., item difficulty and discrimination, systematic and random irregularities) and some remained controversial (e.g., test constructs). In the following, I will discuss the evidence, concerns, and controversies related to the fairness of the GSEEE policies and practices.

Item Quality

Item quality emerged as a common theme among the test takers. It is the most frequently reported theme, indicating the importance of the technical properties in

---

17 An informal interview was conducted with an official from the provincial NEEA branch when I was in the field for data collection.
discussing fairness among the test takers. In general, the discussion related to item quality points out two issues. First, the results regarding item quality generally show that the GSEEE may not provide test takers with sufficient, appropriate opportunities to demonstrate their English proficiency. These concerns were consistent with previous studies that examined the quality of the GSEEE based on statistical analyses. Xiao and Zhu (1999) analyzed the item quality of the 1999 administration of the GSEEE; their results found that a half of the dichotomously scored items (45 out of 90 items) had low discrimination values. Study 1 (Song, Cheng, & Klinger, in preparation) in Chapter 3 also found that discrimination values in the 2009 administration of the GSEEE were generally low (< .20 among 29 out of 45 items), and two items had negative item discrimination values. Overall, these results show that, from both the psychometric evidence and the stakeholders’ perceptions, the GSEEE items do not function well as a fair means to differentiate high and low performers. The results indicate limitations in the GSEEE item design and quality control problems.

Second, when the test takers discussed their perceptions of test fairness in the GSEEE test design and development, their concerns are different from what was defined in the operational model — test bias for test taker groups (1.1 in Figure 2). The test takers who expressed their concerns on item quality did not belong to a specific group of gender or academic background. The result is consistent with Fox’s (2003) study that found it difficult to define group in bias research. In Fox’s study, those who expressed satisfaction or dissatisfaction with the topic were “not of a particular sex, linguistic, or cultural group; they did not follow a particular domain of study, were not exclusively graduate or undergraduate students, nor were they distinguishable by their level of language
proficiency alone” (p. 42). Such result shows how item quality is related to test fairness for individual test takers. Unfairness may be linked to any systematic errors due to the influence of construct irrelevant variables, beyond bias at the group level. This result is important because fairness investigations in test design and development tend to focus on test taker groups.

*Standardized Administration*

Results found that considerable measures were used to ensure the security of the GSEEE administration; however, systematic and random irregularities still occurred in the GSEEE, with random irregularities being cited more frequently. Similar results have also been found in other high-stakes testing in China (Huang & Garner, 2009). The occurrence of irregularities may be related to the high stakes of the GSEEE, which determines test takers’ life opportunity for higher education. The motivation to raise the test scores through cheating becomes too tempting for some individuals to resist, in the past (Yu & Suen, 2005) and at present (e.g., the GSEEE). Standardized administration continues to be a challenge to ensure that all test takers are tested against the same criteria and under the same conditions at all times and locations. Test accommodations, which have been increasingly accepted and practiced in the literature in the West, are not evident in the GSEEE testing policies and practices. As such, stakeholders’ perceptions on this aspect are unclear.

*Scoring*

Both fairness evidence and concerns exist in scoring. While the teachers and test takers welcomed some policies and practices such as rating on the computer, concerns about fairness, particularly about the writing template, emerged. There are at least four
important issues in scoring. First, it is important to eliminate inconsistent scoring practices that lead to overestimating or underestimating the GSEEE test takers’ performance in the writing component (Elder, 2007). Considering the rating problems (e.g., one rater in rating essays), the NEEA needs to provide raters with appropriate training so that construct-irrelevant variables are eliminated and consistency between different raters and different ratings of the same rater are maintained. Second, the NEEA should take initiatives to identify essays using writing templates through systematic analyses. Brooks (2009) examined whether the techniques and computer programs, for example, in lexical overlap, type/token ratio, richness core, and collocations, effectively supported a claim of test takers’ cheating in the translation tests. Such advances would help to identify writing templates and establish defensible criteria for scoring. Third, the impact of writing templates on test scores should be made clear, not only to raters but also to the public. Consistent criteria are needed for raters to follow in order to maintain reliability. Such consistent criteria are also beneficial to the public, particularly test takers, coaching programs, and testing material publishers, so they will adjust teaching, learning, or testing strategies. Fourth, good practices such as rating on the computer should be encouraged across the whole country. It is important to continue to ensure the credibility of test scores by implementing those testing policies and practices.

Cut-scores and Admission Decisions

The last theme, cut-scores and admission decisions, largely allows us to understand how the GSEEE is used for the selection purpose, providing no information on score prediction and predictive bias. The GSEEE seems to be disconnected from learning and teaching at the master’s programs. The participants’ perceptions of the
fairness of the GSEEE policies and practices in decision-making reflect three points. First, the fairness in the GSEEE cut-scores as perceived by the participants can be defined as highlighting equality with consideration of equity (e.g., ethnic minorities). Their conceptualization of test fairness is consistent with the GSEEE testing policies, guidelines, and practices. Such conceptualization of the fairness of decision-making in the GSEEE is different from what is recommended in 1999 Standards, which is promoted in the United States (AERA, APA, & NCME, 1999). The 1999 Standards stresses that the fairness in decision-making can be achieved if test takers who perform equally well on the criteria measure would have an equal chance of being chosen, regardless of group membership. The differences between the stakeholders in the GSEEE and the 1999 Standards may be derived from deep philosophical and ideological value judgments over how test fairness can be pursued, and for whom. Different contexts adopt different testing policies, guidelines, and practices, and stakeholders in different contexts have different value judgments about the fairness in decision-making (Willingham & Cole, 1997; Camilli, 2013).

Second, despite the multiple criteria for decision-making, the question remains whether these measures in two rounds, predominantly testing, truly provide an accurate and holistic estimate of English proficiency. As reiterated in the literature, testing tends to narrowly define proficiency and ability and bring negative consequences on learning (Alon & Tienda, 2007; Zwick, 2002). However, testing (e.g., the GSEEE) still plays a prominent, major role in deciding who can enter master’s programs in China. Although various measures are used for the admission of master’s programs, the challenge remains whether these testing-centred measures provide an accurate, comprehensive estimate of
test takers’ proficiency. In addition, the physical examination is one criterion for the selection purpose. The selection system appears to be organized in such way that test takers’ cognitive ability (narrowly defined by testing) and physical fitness are used to determine if scarce resources should be invested in training them in higher education.

Third, the results show that a process of decentralization has started, for example, giving the 34 universities the autonomy to independently decide their cut-scores since 2004 (Wang, 2009). Decentralization, which allows the consideration of differential needs and expectations at the provincial and institutional level, often implies individualized treatment by case basis (Hanson, 1998). However, the process of decentralization is limited. Although universities have received unprecedented autonomy regarding who can enter their programs, the total number of those who can be accepted into programs is related to the quota system that is decided by the MOE. It is unclear whether such decentralization is a complete revolution or more of a delegation of authority with a limited transfer of decision-making.

*Overall Perceptions of the Fairness of the GSEEE*

Despite mixed perceptions regarding the fairness of the GSEEE policies, guidelines, and practices, the three groups of the GSEEE stakeholders perceived that the GSEEE was a fair test. Such results provide direct evidence that the GSEEE enjoys recognition and acceptance in Chinese society as a fair means to opportunities and success. The GSEEE was seen as fair for several reasons in two major areas: the fair testing process as well as the shared beliefs and priorities among the participants in the Chinese testing context.
First, test fairness in the GSEEE predominantly focuses on equality, with minor considerations of equity (e.g., in decision-making for ethnic minorities). In the GSEEE, equality is given much emphasis, and equity is only addressed on a limited scale. The primary principle of equality in defining the fairness of the GSEEE differs from the literature, which generally agrees that test takers take the same or equivalent tests, under the same or accommodating conditions, and their performances are evaluated under the same scoring procedures (AERA, APA, & NCME, 1999; Xi, 2010). This principle of equality remains the same with the key feature of the Imperial Examinations; that is, all test takers take the same test under the same conditions with the same scoring procedure. Further, the stakeholders’ perceptions on the conceptualization of the fairness of the GSEEE are consistent with the current GSEEE testing policies and guidelines. This may be related to the cultural pattern of the overall Chinese collectivist society in which conformity and obedience to authority is highly valued (Huang & Harris, 1973).

Second, the GSEEE is viewed as fair also due to the situation—the long-standing influence of various social, cultural, and economic considerations that make alignment with the stakeholders’ priority of equality. The GSEEE primarily serves the social need for the selection purpose within the Chinese context, which traditionally values competitive exams. From the perspectives of these participants, the fairness implies a testing-centered, merit-focused system so that students with good testing performance will reach the top and be selected. Testing itself is considered one of the prime virtues and main values in Chinese society. It offers a level playing field for all test takers, rich, poor, from a powerful or weak family background. The testing-centered, merit-focused system, a part of the Chinese educational and philosophical heritage, is one of the major
values that the Chinese government and its stakeholders accept and embrace. In addition, the GSEEE is administered under various human and material constraints. Considering resource limitations and practicality, the one-shot GSEEE focusing on equal treatment has tremendous administrative appeal. The issue of practicality has been discussed in the literature (Backman & Palmer, 1996; Moses & Nanna, 2007). The beliefs focusing on practicality and the social purpose of selection provide an example that high-stakes testing has little, if anything, to do with educational values. Overall, how the issues of the tradition, meritocracy, and efficiency as perceived by the stakeholders can compromise fairness in the GSEEE that influences the well-being of millions of test takers certainly merits future investigation.

To conclude, the study indicates that the fairness of the GSEEE, from the perspective of these participants, should focus on prioritizing equality with consideration of equity on a limited scale. The primary focus of equality in the conceptualization of the GSEEE aligns with the socio-cultural considerations as expressed by the participants in this study. As such, the GSEEE receives overall acceptance and recognition. Whether a test is fair or not is related to testing policies and practices in terms of what to test, how to test, how to administer and score, and how decisions are made at the operational level, with various social, cultural, and practical considerations.

Conclusion and Implications

In the current study, the three groups of the GSEEE stakeholders who had different knowledge about the GSEEE offered information related to the fairness of the GSEEE. The interview results showed evidence, concerns, and controversies regarding the fairness of the testing policies, guidelines, and practices. Regardless, the GSEEE was
considered to be fair due to the reasons within its testing process as well as the participants’ socio-cultural considerations. Overall, the accounts of the GSEEE test stakeholders provide alternative lens to identify concerns about fairness that may otherwise remain undetected by traditional psychometric approaches such as DIF. These issues will, hopefully, attract attention from the GSEEE policy makers and test developers and will promote further data-driven exploration and empirical investigations. The NEEA needs to produce high quality items, establish defensible criteria for scoring essays that use writing templates, and continue to improve test security and minimize cheating. Further, this study provides information on how test fairness, besides its key features within the testing process, is also rooted in contextualized traditions and beliefs. These foundational reasons, going beyond the features in the testing process, contribute to the acceptance of the GSEEE among the interviewed participants. Stakeholders’ perceptions portray a large picture that connects policies, practices, and cultural beliefs in discussing test fairness.

It must be noted that this was a small-scale, qualitative study, focusing on only one test in one selection system. It is possible that socio-cultural factors such as conformity (Hofstede, 2001) or being afraid of “losing face” (Qi, 2011) in the Chinese context affect the trustworthiness of the interview and focus group data. Moreover, the participant sample was only a small fraction of the total number of the GSEEE stakeholders. Perceptions may vary when using different samples and different stakeholders such as test developers or other subgroups of test users. While the participants’ criteria were established to ensure that meaningful information would be attained, the results are only representative of the perceptions of the study participants;
there are limits to generalizing these results to other GSEEE test stakeholders. Given the sampling limitation associated with the qualitative method, there is a need to continue to engage other stakeholders in contributing additional fairness evidence. Although this study suggests that systematically eliciting test users’ and takers’ perceptions helps to pinpoint how to make a test fairer, the results certainly need further investigation and should be incorporated with psychometric evidence. How social, cultural, and political considerations shape the complexity and dynamics of test fairness also needs more focused and thorough investigations.
CHAPTER FIVE
TEST FAIRNESS OF THE GSEEE

In the final chapter, I offer an overall discussion of test fairness as explored in this doctoral research. The chapter combines the findings from the review of the literature and of the two empirical studies and describes how these contribute to the understanding of test fairness of the GSEEE. I first discuss how the empirical studies addressed the two general questions outlined in Chapter 1 through the lens of the conceptual framework adopted in this dissertation (Willingham & Cole, 1997; Willingham, 1999). Then I discuss the implications of this dissertation for the GSEEE’s practices and policies and for fairness theory. Finally, I acknowledge the limitations of the dissertation and suggest avenues for future research on test fairness.

Revisiting Willingham’s Conceptual Framework

As described in Chapter 2, Willingham conceptualizes test fairness as centred by comparability investigations for groups and individuals of test takers across multiple aspects of the testing process involving stakeholder groups within a testing situation (Willingham & Cole, 1997, Willingham, 1999). Fundamentally, his conceptual framework consists of four elements: comparability, process, participants, and situation. Through the lens of this conceptual framework, I discuss the major findings of my dissertation in the following.

Comparability

The central principle of test fairness in Willingham’s work is comparability – the comparable treatment for test takers, both groups and individuals. Willingham uses comparability to represent multiple dimensions of fairness such as absence of bias,
equality, and equity. In Study 1, I examined item bias for test taker groups. Using SIBTEST and content analysis, I investigated whether the GSEEE items functioned differentially and brought advantages or disadvantages towards certain test taker groups (e.g., gender and academic background). In Study 2, I investigated stakeholders’ perceptions of the fairness of the GSEEE towards groups and individuals of test takers across the testing process. Overall, the results of Study 1 and 2 did not find evidence to suggest that the GSEEE test items showed potential bias between males and females or across different academic backgrounds (Humanities & Social Sciences or Sciences). Further, based on the results of Study 2, test fairness in the GSEEE focuses on equality, with minor considerations of equity toward certain disadvantaged groups in score-based use (e.g., lowered cut-scores for ethnic minorities).

Equality, as described in Chapter 2, means that test takers are treated as identically as possible, regardless of their group membership (Caldwell, Shapiro, & Gross, 2007). Equality connotes sameness and disregards treatment differences. Equality was found to be a major principle in the GSEEE in that test takers were required to take the same test at the same time, be treated identically in the test administration, and be scored impartially without favouritism. Admission decisions were primarily made based on the GSEEE score ranking. In general, equality is the central feature of various testing systems in China, which includes the GSEEE. It has been promoted, practiced, and perceived as a major principle to ensure fairness since the Chinese Imperial Examinations (606 AD-1904AD) for many years. Such a conceptualization of test fairness focusing on equality does not seem to have changed within the current Chinese context. Compared to the drastic evolution of the concept of test fairness as demonstrated in the literature, the
primary feature of test fairness of the GSEEE remains the same. The findings of the dissertation confirm that equality continues to drive stakeholders’ perceptions and the GSEEE testing policies and practices to achieve its fairness claims.

Equity, on the other hand, requires differential and special treatment under certain situations (Shapiro & Stefkovich, 2001). Equity, in comparison with equality, was addressed on a limited scale in the GSEEE, only in relation to score-based use. Two groups were given equitable treatment in the GSEEE policies: ethnic minorities and those who planned to pursue their master’s study in economically less-developed areas (often populated with ethnic minorities). These two groups were compensated through lowered cut-scores, due to the consideration of unbalanced learning opportunities. The GSEEE administrators, teachers, and test takers endorsed such equity policies, guidelines, and practices favouring the two disadvantaged groups. In current China, while espousing the major principle of equality, policy makers such as the MOE recognize this principle as limiting and problematic, particularly in situations where there are unequal distributions of educational opportunities for certain groups (Liu, 2009; Wang, 2007). As such, in the GSEEE, special attention is given to fairness for the two groups, ethnic minorities and those from the western less-developed areas, in decision-making. Equity thus is practiced, focusing especially on accelerating the admission rate of the two disadvantaged groups through lowered cut-scores.

In general, the stakeholders’ perceptions on the conceptualization of the fairness of the GSEEE are consistent with the features of the current GSEEE testing policies, guidelines, and practices, that is, equality with minor considerations of equity. The stakeholders’ conceptualization makes alignment with the testing policies and practices
developed by the MOE and NEEA. Such consistency may be derived from the Chinese collectivist culture where conformity and obedience to authority is highly valued (Hofstede, 2001; Huang & Harris, 1973). Collectivism encourages conformity among the GSEEE test takers and users to accept and endorse the testing policies and practices stipulated by the MOE and NEEA in achieving the goal of selection across the whole country.

**Process**

The second component of Willingham’s conceptual framework focuses on the testing process, which ranges from test design and development to score-based use. Willingham’s work highlights the investigation of test properties and examines activities and practices directly related to testing, though not necessarily beyond the testing process (e.g., opportunity to learn). In Study 1, I examined one aspect of the testing process, that is, biased items in test design and development. In Study 2, I investigated fairness perceptions across the testing process from test design and development, administration, scoring, to score-based use. The results of Study 1 and 2 together identified fairness concerns. Study 1 found the reliability estimates lower than the 0.70 standard (Pedhazur & Schmelkin, 1991) and low discrimination values (29 out of 45 MC items below .20) of the GSEEE test items. Two items had negative discrimination values. Study 2 discovered low item quality as perceived by the test takers, administrative irregularities (e.g., cheating), and inconsistent scoring practices regarding how to identify essays that used templates as well as how to score them as perceived by the teachers and test takers.

Overall, these results indicate serious issues relating to test quality, in particular, test reliability. Reliability, one of the most important elements of test quality, refers to the
consistency of a measure. If a test is designed to measure a trait (e.g., English proficiency), test scores should be approximately the same when the test is administered to a test taker in different administration sites or assessed by different raters. Pedhazur and Schmelkin (1991) believe that “very high reliabilities are essential when the scores are used for making important decision about individuals” (p. 109). Given the current empirical evidence (e.g., the low internal reliability estimates, administrative irregularities, and the inconsistent scoring practices), the GSEEE as a norm-referenced test does not provide reliable results to estimate test takers’ English proficiency through its test items, across different administration sites, or based on its scoring procedures. The GSEEE test takers with similar English proficiency could receive different scores if they sit in the different testing centres or scored by different raters. Such results threaten the fairness claim of the GSEEE because some test takers are not provided with equal opportunities to demonstrate their ability in comparison with other test takers. The GSEEE potentially underestimates (or overestimates) the competence of test takers who would have done better (or poorer).

Participants

Willingham states that test participants should be incorporated into the discussion of test fairness because of their involvement in the testing process. While the views from test takers as well as professionals and institutions may conflict, they all “have a legitimate claim” due to their involvement (1999, p. 224). In Study 1, I analyzed item-level data and invited three content reviewers who were also item writers, in the GSEEE or other Chinese high-stakes testing, to examine the potential bias of test items. In Study 2, perceptions of the fairness of the GSEEE were collected through one-on-one
interviews and focus groups from program administrators, English teachers, and test takers. Overall, the item writers in Study 1 provided limited insights as to what caused DIF. They believed that the existence of DIF was due to some anecdotal reasons (e.g., female test takers tended to be more motivated and diligent than males). The item writers did not go beyond stereotyping input. In Study 2, bias did not emerge as a predominant theme from the interview data. Different groups of test stakeholders had different levels of knowledge about the GSEEE. While the test takers provided information related to all the six aspects of the GSEEE process, the program administrators and teachers provided information only in select aspects (e.g., score-based use). The administrators and teachers lacked information about test design and development, and administration.

On the whole, despite identifying various issues related to the fairness of the GSEEE policies, guidelines, and practices, the item writers, program administrators, teachers, and test takers generally did not provide in-depth, thorough, and comprehensive information about the conceptualization of test fairness. Their views were only at a simplistic level. They did not appear to understand or show awareness of some key dimensions of test fairness, such as lack of bias and equity for test takers with special needs, which have been substantively discussed in the literature (Abdi, 2007). Further, Willingham and Cole (1997) argue that different stakeholder groups may “see issues through different lenses” and “suggest different priorities for fair assessment” (p. 365); however, the administrators, teachers, and test takers in this study shared their views and priorities, that is, prioritizing equality with minor considerations of equity toward the two disadvantaged groups in score-based use. Again, such consistency across different groups of stakeholders on the conceptualization of the fairness of the GSEEE may be related to
the Chinese collectivist culture. Given the Chinese culture where conformity is highly valued, it is difficult to obtain insightful accounts among test takers and users regarding the conceptualization of test fairness.

The finding that different groups of test stakeholders had different levels of knowledge coincides with Willingham’s framework that stakeholders “have different connections with the process” (1999, p. 224). While the test takers were widely involved in all the four areas of the testing process, the administrators and teachers had relatively weak connections with the GSEEE. Such results come as no surprise because the three groups have different experiences and interests. The test takers are familiar with the GSEEE through their test-taking experience. In comparison, the program administrators and teachers are less interested in test design and development, and administration, partially because the GSEEE is mandatory and test users are not in a position to select another English test. They have limited influence on testing policies, guidelines, and practices; hence they are not actively involved in all the aspects of the testing process.

Situation

Willingham recognizes the role of the situation, in particular its influence on the fairness of one area of the testing process, score-based use. He states that the testing situation can vary in numerous ways – “the purpose to which the test is put, the side effects of its use, additional variables that are used along with the test, the criterion that is used, the particular context of use, and the specific sample that is involved” (1999, p. 224). Fairness is mediated by a particular testing situation with various social, cultural, economic, political, and legal criteria and considerations. The situation that the GSEEE operates in has its unique social and cultural features, as elaborated largely by the
stakeholders in Study 2 of this dissertation. These socio-cultural considerations that contributed to the stakeholders’ perceptions of the fairness of the GSEEE include: (1) the tradition of testing, (2) the testing-centred meritocracy, and (3) pursuit of efficiency. In this section, I elaborate on the three socio-cultural mediators that are embedded in the testing situation as expressed by those who participated in Study 2.

The Tradition of Testing

[We] Have used testing for so many years, for entering high school, university, master programs, or doctoral programs. [We are] used to them, feeling like a part of our life. (Test taker)

The three groups of the participants believed that the GSEEE was a fair test, partially due to the testing tradition in China. As stated earlier, testing can be traced back to 606 AD when the emperor in the Sui Dynasty (581 AD-618 AD) decreed that exams were the official criterion for those who wished to serve in government. Although the Imperial Examinations were officially abolished in 1905, the strong influence of thousands of years of the testing tradition in Chinese education remains, and examinations continue to be used as a major tool in day-to-day classroom teaching and learning (Berry, 2011). Paper-and-pencil examinations, which highlight equality in the testing process in China, have been widely accepted as the influential, legitimate indicator of proficiency, achievement, and promotion in every corner of the current society including school and university admission, professional certification, selection for employment, and job promotion (Zhang, 2007). High-stakes testing as an important fact of life has been imprinted in every step of students’ lives.
Chinese society and its people have been accustomed to examinations and have culturally accepted high-stakes testing as a means to determine their future prospects and learning opportunities for thousands of years (Yu & Suen, 2005). Testing has become a common standard and begins so early in a person’s life that few people question its legitimacy. The tradition has become deeply entrenched in Chinese culture so that testing itself is highly valued. Hence, the GSEEE is seen as a symbolic representation of knowledge and skills that are required for the admission to master’s programs. It has tremendous appeal to its stakeholders who are accustomed to, and even obsessed with the numerical measurement of English proficiency. The fairness of the GSEEE is thus taken at face value.

The harmful impact of the test-oriented culture has been found in educational research (Wang, 2008; Zhang, 2007). Various national initiatives and reforms have been adopted to address problems related to the deeply entrenched testing-oriented system since the 1990s (Yang, 1993). For example, the recent curriculum standards drafted by the MOE stress that the country’s assessment system should change from being dominantly reliant on the selection function of assessment to its formative use for enhancing learning and supporting teaching (Ministry of Education, 2003). However, these policies are not detailed, nor they are consistent. The curriculum documents fail to provide concrete suggestions and strategies to realize a true shift away from the test-oriented tradition in its educational system (Nisbett, 2003). High-stakes testing continues to be used as a major tool to decide admission, make judgments of learning outcomes, and drive teaching and learning (Berry, 2011).

*The Testing-centred Meritocracy*
How could you give those poor [economically deprived] kids an opportunity to compete against guan er’dai and fu er’dai if testing is cancelled? (Administrator)

The Chinese term guan er’dai (官二代) can be translated into English in numerous ways such as “sons and daughters of government officials” or “official offspring.” This Chinese term is used with increasing frequency to refer to the children of current or former Communist Party or government officials with privileges and opportunities by virtue of the power enjoyed by their parents or grandparents. The term is used in conjunction with the term fu er’dai (富二代), or “progeny of prosperity,” which refers to the sons and daughters of rich, powerful businessmen, who are similarly afforded tremendous opportunities. Both terms are widely used in the current Chinese context, implying growing inequality due to the differences in educational opportunities, social connections, and economic status. To combat such inequality, as was promoted in the Imperial Examinations, testing such as the GSEEE has been used to provide a more level playing field for test takers to demonstrate their merits. The fairness in score-based use would be achieved through selecting individuals who have high scores, regardless of their group membership.

Meritocracy highlights that opportunities are assigned to individuals based upon “merits”, namely intelligence and ability, commonly determined through examinations or evaluations (Young, 1958). Meritocracy, as symbolized by equality in testing and open competition (Young, 1958), is a valued ideal in Chinese society (Yao, 2007). It embraces a system where individuals, regardless of their background, are rewarded based on their test performance and effort. Although the test takers interviewed in Study 2 did not always agree on the specific constructs to be tested in the GSEEE as merit (e.g., whether
listening and speaking should be included in the GSEEE), the testing-focused meritocracy is a primary consideration. However, when seeking a “level playing field” through testing, we often see that the playing field in learning opportunities is not level. Students from middle-upper economic status tend to perform better than those from low economic status in testing (Zwick, 2012). Meritocracy that is based on equality in the testing process actually does not eliminate the advantages afforded by socio-economic wealth and family background. As such, the fairness of the GSEEE also considers preferential treatment for two groups in the GSEEE score-based use, given their unbalanced learning opportunities. Regardless, the testing-centred meritocracy focusing on equality in the testing process is predominantly valued in the Chinese context (Yao, 2007). Overall, in this context, the GSEEE is perceived as fair because the society is organized in such way that test takers’ abilities, which are narrowly defined by test scores with limited considerations of differences in learning opportunities, are used to determine if resources should be invested in training them for certain desirable educational opportunities.

Pursuit of Efficiency

_We have about 3000 spaces in our university, and it takes forever [to complete the selection of qualified candidates for our master’s programs] so we have to use the GSEEE._ (Administrator)

The overall perception that the GSEEE was a fair test was also mediated by the two socio-cultural considerations: (1) usefulness in terms of the GSEEE’s major goal of selection under the national agenda and (2) practicality that considers financial constraints and meets the needs of efficiency in selection. These two considerations can
be summarized as the pursuit of efficiency, which mediates the stakeholders’ overall perception of the fairness of the GSEEE. Currently, the national high-stakes GSEEE is administered to millions of test takers every year. The GSEEE, based on the principle of equality, provides a “quick and easy” solution to the goal for admission purposes. The current test design of the GSEEE, having eliminated listening and speaking, meets the administrative challenge of serving millions of test takers simultaneously. Further, given the present financial limitations, the one-time GSEEE, focusing on equality, has tremendous administrative appeal and is cost-effective as an efficient method for screening applications at the national level. The GSEEE scores provide a proxy for merit, and are easily compared between test takers. Moreover, the GSEEE serves purposes beyond admission decisions at individual institutes. As an example, the GSEEE results are used to place students or choose which students can waive English courses. These uses of the GSEEE scores are adopted simply because they provide a great deal of efficiency, practicality, and administrative utility.

The issue of usefulness, practicality, and efficiency has been discussed in the research literature (Phelps, 2005; Walberg, 2011). Willingham (1999) believes that “the social justification of a test sits on the three-legged stool of fairness, usefulness, practicality” (p. 227). As argued in *Principles for Good Practices by ALTE Examinations* (ALTE, 2001), fair test practices should be supported within the limits of feasibility and acceptability. It points out that all tests are context specific and therefore “practical considerations and constraints have to be considered through all stages of the testing process” (p. 15). The authors state that it is imperative to determine whether a test is feasible, can be produced and administered with the available resources, and at a
manageable cost. Decisions on these issues should be supported by reasonable economic, social, and educational rationales. Bachman and Palmer (2010) also contend that a fair test is a process of trade-offs within a testing situation. They believe that fair tests may have to reduce “the importance of one or more qualities of a claim in order to maintain or increase the qualities of another claim, either in response to competing values of different stakeholders, or in order to make the assessment practical” (p. 266). Alternatively, fairness may be compromised by practicality. Obviously, there are not universal trade-offs that can be applied in all testing situations. In the GSEEE, the testing practice seems to prioritize usefulness, practicality, and efficiency at the cost of fairness (e.g., construct under-representation due to the exclusion of listening and speaking). How the trade-off influences the GSEEE test takers’ opportunities to demonstrate their knowledge and skills at their best certainly deserves further investigation.

To conclude, the tradition of testing, meritocracy, and the pursuit of efficiency embedded in the situation that the GSEEE operates in were identified as the key considerations mediating the stakeholders’ perceptions. These three socio-cultural considerations make alignment with the stakeholders’ perception of prioritizing equality, which results in the acceptance and endorsement of the GSEEE as fair. The GSEEE is perceived as a reasonable and meaningful representation of knowledge and skills for the admission purpose, just like many other culturally specific practices. It becomes a powerful symbolic representation that is given epistemic endorsement. Such socio-cultural acceptance among test users and takers may breed and support other forms of acceptance (e.g., political and ideological) in society. These various forms of acceptance strengthen legitimacy of existing equality-prioritized testing policies and practices
promoted by the centralized Chinese government. Test users’ and takers’ endorsement as well as the legitimizing forces of policy makers and institutionalized policies combine together to perpetuate the existence of the GSEE. From this perspective, the GSEE that is seen as fair within this Chinese situation is justified to the extent that it is connected to an underlying symbolic socio-cultural system, regardless of the fairness concerns in the testing process.

Implications

The implications of each of the two studies have already been discussed in Chapters 3 and 4 separately. In this section, I discuss the implications of the dissertation as a whole for the GSEE’s practices and policies as well as fairness theory.

Implications for Practice

This dissertation provides important implications regarding GSEE testing practices, including the urgency to improve test quality, increase test developers’ openness in communicating quality evidence to the public, and develop public understanding of test fairness. First, although the NEEA claims to have established a quality control system and conducted test evaluation research (Liu, 2010), both Study 1 and 2 identified quality issues in the GSEE, which weaken the GSEE’s fairness claim. As discussed in the section of implications in Chapter 3 and 4, it is imperative for the NEEA to improve test quality of the GSEE from various aspects, for example, producing high quality test items, minimizing cheating, and establishing defensible criteria for scoring essays using templates.

Second, it is critical that the NEEA develops openness in communicating research evidence to the public. The whole testing system in China is rather closed with little
transparency. Although the NEEA claims to have conducted research activities (Liu, 2010), it is unclear if a proper communication mechanism about research results has been established. Test evaluation reports are not available to the research community and the public. The NEEA officials also showed little willingness to engage in communication and interaction when they were approached. I had an opportunity to meet with an official from the NEEA at an international conference that was held in North America. When I expressed my interests in reading the GSEEE test evaluation reports and talked about my potential interview with him, the official declined, replying “not possible.” During the trip to conduct my pilot study, I was unable to find anyone from the NEEA who agreed to be interviewed because they were continuously “not available for an interview” [不便采访]. After trying all the different sources and connections available to me as an external researcher, I finally found an official from a provincial Examination Board who agreed to be interviewed. However, just before the interview, I was told that I could not record the interview due to the sensitive information being discussed and the official’s confidentiality obligations to the NEEA. This is not surprising due to the large-scale and high-stakes nature of the GSEEE – one test for all test takers in the big country of socialist China. Within the Chinese Communist Party-led political system (Lawrence & Martin, 2012; Liu, 2001), the official who represented the NEEA and MOE had to be protective about any information related to the GSEEE to be disclosed to the research community and public because such information might bring negative influence on the Communist Party’s leadership.

Third, the dissertation provides implications that more actions and education ought to be taken to increase public awareness regarding what constitutes the fairness of a
test (Hamp-Lyons, 2000). These actions may include, but are not limited to: use of media in discussing test fairness, legal rights and responsibilities and associated consequences, organizing awareness raising campaigns, increasing education in target groups, and encouraging public participation in testing matters. Capacity building efforts could provide the public with information and knowledge about test fairness. Education, training, and awareness programs for program administrators, teachers, and other professionals could help to enhance stakeholders’ understandings of key elements related to test fairness such as bias and equity. Most importantly, test users and takers should be encouraged to be involved in issues related to test fairness. As Rea-Dickins (1997) states, stakeholder participation is “not only seen as providing a forum for stakeholder concerns, ideas, complaints and opportunities for affirmation of decision made, but it is also about equipping teachers, parents and others with information so that they may take appropriate action” (p. 311). It is important to establish a notion of shared responsibility—the responsibility for test fairness is in the hands of all those who are involved in the testing.

Implications for Policy

This dissertation also provides significant implications for the Chinese Ministry of Education (MOE) in terms of its testing policies in the GSEE. As shown in the results of the study, much of the discussion regarding the fairness of the GSEE focused on equality. Equity was only addressed in score-based use with two disadvantaged test taker groups. Other distinctive groups and individuals such as physically disabled test takers, 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; Wang, 2007), are largely neglected. This provides policy implications. Services
such as test modifications and accommodations as well as educational resources and learning opportunities need to be provided so that these disadvantaged test takers have comparable opportunities to demonstrate their knowledge and skills. In spite of being widely emphasized in other contexts (e.g., Canadian and U.S. testing programs), test takers with disabilities and learning difficulties remain the most vulnerable and invisible group in China (Human Rights Watch, 2013). Accommodation practices are not provided in any large-scale high-stakes testing in China, including the GSEEE (Fan & Jing, 2012).

Similarly, the use of physical fitness condition as one criterion for admission to master’s programs results in significant barriers for disabled test takers. Based on the national ministry policies and guidelines, all test takers applying for universities must submit the results of physical examinations (MOE, 2003). Those who have “physiological defects” and “illnesses” that make a person “unable to take care of themselves or complete their studies” can be grounds for denying their admission to universities (MOE, 2003, p. 5). Hence disabled students often lack educational accessibility and learning opportunities. There is still a great deal to be accomplished in order to truly realize participation, inclusivity, and diversity in China for those who are in a disadvantaged position, including disabled test takers.

**Implications for Theory**

This dissertation is guided by Willingham’s framework (Willingham & Cole, 1997; Willingham, 1999) and by the research conducted in North America. I adopted Willingham’s conceptual framework because his framework provided an elaborative description about test fairness (Welch, 1998). As a researcher working for ETS then, Willingham built his career on a variety of educational topics, in particular, handicapped
test takers (Willingham, 1988) and college admission (Willingham, 1985) in the context of the United States. Willingham’s framework has been situated in a society that highlights social justice, especially for disadvantaged groups (Camilli, 2007; Gordon & Bonilla-Bowman, 1999). Willingham’s work, particularly his research on handicapped test takers, is influenced by distributive justice theorists such as John Rawls—a well known, leading philosopher in the United States (Zajda & Majhanivich, 2005). Rawls (1971) proposes a social justice framework that includes two equally important principles: justice requires not only (1) equality in the treatment of all members of the society but also (2) the protection of the least advantaged members of the society in order to make them contributing citizens. Situated in the context of the United States and influenced by Rawls (1971), Willingham assumes that, in addition to equal treatment among test takers in the testing process, attempts to remove obstacles for disadvantaged test takers (e.g., handicapped test takers) should be equally made. Resources and services such as test modifications and accommodations should be distributed equitably to aid these weakest members of society so their performance can be accurately estimated.

However, the results of this dissertation show that, within the Chinese context where the GSEEE operates in, the stakeholders’ conceptualizations of the fairness of the GSEEE differ from what Rawls proposed and Willingham stated within the testing context of the United States in the 90s. The participants of my study attached much attention to equality with minor considerations of equity, and their priority to equality was mediated by socio-cultural conditioning in which the tradition, meritocracy, and efficiency were highlighted. Because the GSEEE stakeholders are situated and ingrained in the Chinese context thus bearing different values from those in the United States, their
views of test fairness are inevitably different from Willingham’s stance. These results provide implications for fairness theories and suggest the importance to emphasize the differential role of the *situation* plays in using Willingham’s framework. There is a room to develop Willingham’s fairness conceptual framework and add more weight to the element of the *situation* because it mediates the conceptualization of test fairness. In addition to the testing process, future fairness studies should investigate how various contextual forces (e.g., political and legal forces) influence test fairness as well as how they are related to each other and their relative strength. The study identified three socio-cultural considerations. Given the mediating role of the situation in conceptualizing test fairness of the GSEEE, future studies need to identify other contextual considerations. Through inter-disciplinary collaborations with researchers in different fields (e.g., political science and law) (Kunnan, 2008; Moss, et al., 2005), it is possible to have in-depth, thorough examinations of how various social, cultural, economic, political, legal, and philosophical forces that are embedded in the Chinese context mediate the current conceptualization of test fairness.

The results of this dissertation also imply that, in the use of Willingham’s framework, the scope within the testing process for fairness investigations should be expanded, for example, to include opportunity to learn. The dissertation research is an initial attempt at adopting a framework for fairness investigations that was rooted in Willingham’s work, primarily focusing on the testing process (Willingham & Cole, 1997; Willingham, 1999). Although much information related to the fairness of the GSEEE in the process was found, some discussions fell beyond the testing process, which is not part of Willingham’s framework. The participants stated that unbalanced learning
opportunities led them to conclude that it was fair to use preferential treatment in decision-making for the two disadvantaged groups. The perceptions of the fairness in decision-making for these groups appear to be linked with the unfairness beyond the testing process. Unequal learning opportunities impact testing policies, practices, and perceptions. Fairness investigations that are defined as focusing on the testing process, according to Willingham’s work, are actually related to issues outside the testing process. This provides evidence that the scope of fairness investigations may need expansion to include areas such as opportunity to learn as defined by Gipps and Stobart (1999) and Kunnan (2008). Further research needs to refine Willingham’s framework or consider fairness theories that can explain how testing policies, practices, and perceptions are influenced by disparities in educational conditions among different groups.

Limitations and Future Research Directions

I discussed the limitations of the two studies separately in Chapters 3 and 4. Here, I identify limitations of the dissertation as a whole. First, the two studies focused on different administrations of the GSEEE. Study 1 examined test items from the 2009 administration. Study 2 investigated stakeholders’ perceptions of the fairness of the 2012 GSEEE administration. These decisions were made due to the sequential nature of the studies and feasibility. A focus on one test administration may have provided more descriptive information regarding test quality and the fairness of specific test items using DIF statistically analysis and from the perspectives of the test stakeholders. Second, test developers are not included in this dissertation due to their unavailability. This important group needs to be included because they have the resources and knowledge to provide more in-depth and precise information regarding testing policies and guidelines,
specifically on test design and development. Third, in Willingham’s book regarding the fairness framework, many aspects of the testing process are pointed out for fairness investigations such as comparability across parallel forms. This doctoral research only focused on limited aspects. How fairness would be demonstrated in other aspects of the GSEEE is unknown.

Considering that test fairness is “an extraordinarily broad subject” (Willingham & Cole, 1997, p. 234), more research is needed to help to identify which aspects of the testing process may threaten fairness, for whom, and in what socio-cultural situations. Based on my research experience on this topic, I have identified several directions for future research. First, further research could include test developers (e.g., officials working in the NEEA and MOE) and examine their views on test fairness because they are key informants who establish testing policies, provide research evidence in support of testing practices, and can best explain the rationales for adopting certain testing policies. Officials working in the NEEA and MOE have access to all the information related to the test and test takers. They take responsibility for selecting, writing, and reviewing test questions, test administration and scoring, and deciding cut-scores. Ideally, they would also possess professional skills and have comprehensive knowledge regarding the science of testing. Although eliciting the perceptions of the GSEEE test developers may not guarantee deeper, insightful views on test fairness, what they can offer is to provide a more detailed, comprehensive description about testing policies and rationales for the use of such policies.

Second, future studies could examine test fairness with other subject tests of the GSEEE and other tests within the Chinese situation (e.g., the National Matriculation
Examination) to fully understand the common conceptualizations of test fairness across various tests in the Chinese context. Fairness inquiries in this dissertation can serve as a generative function by providing a basis for ongoing research to achieve a more refined understanding of test fairness in the Chinese context.

Third, future fairness studies in the Chinese context may adopt a critical testing approach (e.g., Shohamy, 2001). Based on my research, the critical approach can be one of the most promising methodologies for fairness investigations within the Chinese context. Influenced by many philosophers and social theorists (e.g., the Frankfurt School), the critical approach describes how dominant and subordinate groups based on socioeconomic, physical, and gender status struggle over power in ways that make every aspect of life political (Patton, 2002). Shohamy (2001) contends that critical testing “implies the need to develop critical strategies to examine the uses and consequences of tests, to monitor their power, minimize their detrimental force, reveal the misuses, empower the test takers” (p. 131). Critical inquiries do not seek to study and understand issues in a society, but rather to critique and change the society and seek “human emancipation” in circumstances of domination (Horkheimer 1982, p. 244). By empowering certain individuals and groups (e.g., disabled test takers), critical research emphasizes an attempt to confront the injustice of a particular society or public sphere within the society (Kincheloe & McLaren, 2005). This will encourage continued improvement in the genuine legitimacy and acceptance of testing policies, practices, and interpretations.
Conclusion

To conclude, pursuing test fairness is one of the important endeavours in any educational and social context. The responsibility for fair practices and good conduct is in the hands of all those who are involved in testing (Gipps & Stobart, 2009). It is of paramount importance to have continued discussions regarding what constitutes fair treatment to present balanced attention to the interests of groups and individuals of test takers. Great attention to and openness in discussing test fairness will improve tests, offer voice to diverse stakeholders’ interests, and enhance researchers’ capacity to make a difference in testing and education. Possessing one of the world’s largest higher education systems, China merits the attention of educators, policy makers, global scholars, and testing companies. Equally important in meeting the growing demands for higher education, test fairness in China must be achieved, as much as possible, for all test takers and test taker groups across all aspects of the testing process.
REFERENCES


Cobb, T. (1999). Breadth and depth of vocabulary acquisition with hands-on


Helms, J. (2006). Fairness is not validity or cultural bias in racial-group assessment: a


Language Learners on the National Assessment of Educational Progress.

*International Journal of Testing, 8, 14-33.*


Retrieved on May 20, 2012 from


Roach, A. T., Chilungu, E. N., LaSalle, T. P., Talapatra, D., Vignieri, M. J., & Kurz, A. (2009). Opportunities and options for facilitating and evaluating access to the...


Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of


Zumbo, B. D. (2007). Three generation of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223-233.
APPENDICES

Appendix A Ethics Clearance for Study One

April 26, 2010

Dr. Lijing Cheng
Faculty of Education
Duncan McArthur Hall
Queen’s University

GREB Ref #: GEDUC-511-10
Title: “Test Fairness in High-Stakes Testing Decisions: An Investigation of Test-Taker Group Differences”

Dear Dr. Cheng:

The General Research Ethics Board (GREB), by means of a delegated board review, has cleared your proposal entitled “Test Fairness in High-Stakes Testing Decisions: An Investigation of Test-Taker Group Differences” for ethical compliance with the Tri-Council Guidelines (TCPS) and Queen’s ethics policies. In accordance with the Tri-Council Guidelines (article D.1.b) and Senate Terms of Reference (article G), your project has been cleared for one year. At the end of each year, the GREB will ask if your project has been completed and if not, what changes have occurred or will occur in the next year.

You are reminded of your obligation to advise the GREB, with a copy to your unit REB, of any adverse event(s) that occur during this one year period (details available on webpage http://www.queensu.ca/ors/researchethics/GeneralREB/forms.html – Adverse Event Report Form). An adverse event includes, but is not limited to, a complaint, a change or unexpected event that alters the level of risk for the researcher or participants or situation that requires a substantial change in approach to a participant(s). You are also advised that all adverse events must be reported to the GREB within 48 hours.

You are also reminded that all changes that might affect human participants must be cleared by the GREB. For example, you must report changes in study procedures or implementations of new aspects into the study procedures on the Ethics Change Form that can be found at http://www.queensu.ca/ors/researchethics/GeneralREB/forms.html – Research Ethics Change Form. These changes must be sent to the Ethics Coordinator, Gail Irving, at the Office of Research Services or irvingg@queensu.ca prior to implementation. Mrs. Irving will forward your request for protocol changes to the appropriate GREB reviewers and/or the GREB Chair.

On behalf of the General Research Ethics Board, I wish you continued success in your research.

Yours sincerely,

Joan Stevenson, PhD
Professor and Chair
General Research Ethics Board

c.c.: Dr. Malcolm Welch, Chair, Unit REB
Xiaomei Song, Co-Applicant
E-REB: c/o Graduate Studies & Bureau of Research, Attn: Colinas Freitas

JS/II
Appendix B Ethics Clearance for Study Two

January 19, 2012

Ms. Xiaomei Song, Ph.D. Candidate
Faculty of Education
Duncan McArthur Hall
Queen’s University
511 Union Street
Kingston ON K7M 5R7

GREB Ref #: GEDUC-595-11; Romeo # 6006489
Title: "GEDUC-595-11 Test stakeholder perceptions on fairness of a large-scale high-stakes language test"

Dear Ms. Song:

The General Research Ethics Board (GREB), by means of a delegated board review, has cleared your proposal entitled "GEDUC-595-11 Test stakeholder perceptions on fairness of a large-scale high-stakes language test" for ethical compliance with the Tri-Council Guidelines (TCPS) and Queen's ethics policies. In accordance with the Tri-Council Guidelines (article D.1.6) and Senate Terms of Reference (article G), your project has been cleared for one year. At the end of each year, the GREB will ask if your project has been completed and if not, what changes have occurred or will occur in the next year.

You are reminded of your obligation to advise the GREB, with a copy to your unit REB, of any adverse event(s) that occur during this one year period (access this form at https://eservices.queensu.ca/romeo_researcher/ and click Events - GREB Adverse Event Report). An adverse event includes, but is not limited to, a complaint, a change or unexpected event that alters the level of risk for the researcher or participants or situation that requires a substantial change in approach to a participant(s). You are also advised that all adverse events must be reported to the GREB within 48 hours.

You are also reminded that all changes that might affect human participants must be cleared by the GREB. For example you must report changes to the level of risk, applicant characteristics, and implementations of new procedures. To make an amendment, access the application at https://eservices.queensu.ca/romeo_researcher/ and click Events - GREB Amendment to Approved Study Form. These changes will automatically be sent to the Ethics Coordinator, Gail Irving, at the Office of Research Services or irvingg@queensu.ca for further review and clearance by the GREB or GREB Chair.

On behalf of the General Research Ethics Board, I wish you continued success in your research. Yours sincerely,

Joanne Stevenson, Ph.D. Professor and Chair
General Research Ethics Board

cc: Dr. Liying Cheng, Faculty Supervisor
     Dr. Lesly Wade-Woolley, Chair, Unit REB
     Erin Wicklam, c/o Graduate Studies and Bureau of Research
Appendix C Letter of Information for Content Reviewers

“Test Fairness in High-Stakes Testing Decisions: An Investigation of Test taker Group Differences”

I, Dr. Liying Cheng, Faculty of Education, Queen’s University, am inviting you to participate in this study—“Test Fairness in High-Stakes Testing Decisions: An Investigation of Test taker Group Differences”. This study was granted clearance by the General Research Ethics Board for compliance with the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans, and Queen's policies.

The purpose of this proposed study is to investigate group membership effects on the Graduate School Entrance English Examination (GSEEE) with aspects to gender, major of study (Humanities & Social Sciences/Sciences), and university type (key/non-key universities). The study invites the GSEEE test designers to a) decide whether each item of the GSEEE is likely to advantage/disadvantage test takers who are female/male, from Humanities& Social Sciences/Sciences majors, and from key/non-key universities, b) use the 5-Likert scale from strongly disadvantage (1), disadvantage (2), neutral (3), advantage (4), to strongly advantage (5) to rate the suitability of those items, and c) explain their choices in the comment area. In order to do this, you are invited to review contents of items with the GSEEE administered in 2009. I will interview you once via telephone. The interview will take from 45-60 minutes. The interview will be recorded and transcribed.

Participation is entirely voluntary. You are free to withdraw from the study without reason at any point and may request removal of your data. Since the GSEEE test designers (approximately 20) vary each year and the names are highly confidential across a large population of China, there is a minimal risk that your participation will be identified by the public and your comments will have potential influence on your life. However, there is a possibility that your response might be identified by those who already know your opinions. If you decide not to participate, your decision will have no impact on your relationship with the investigator.

Only an identification number will be assigned to your responses. Under no circumstances will your name or your response be released to anyone else beyond the research team. Consent forms, recorded interview, rating responses, and transcription for the interviews will be secured in a locked cabinet in my office until the completion of the study. My research assistant and I are the only ones who will have access to the data. Completed interview transcripts, consent forms, and recorded data will be destroyed at the completion of the study.

This study may result in publications of various types and conference presentations. However, participants’ names will not be attached to any form of the data provided. Pseudonyms will be used in any reporting of the study. An electronic copy of a report from this study will be made available to you upon your request.
Any questions about study participation may be directed to Dr. Liying Cheng, Principal Investigator of the study at 1 (613) 533-6000 ext. 77431 or at liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 1 (613) 533-6081 (chair.GREB@queensu.ca).
Appendix D Consent Form

for

“Test Fairness in High-Stakes Testing Decisions: An Investigation of Test taker Group Differences”

I have read, understood, and retained a copy of the letter of information concerning “Test Fairness in High-Stakes Testing Decisions: An Investigation of Test taker Group Differences”. The purpose of this proposed study is to investigate group membership effects on the Graduate School Entrance English Examination (GSEEE) with aspects to gender, major of study (Humanities & Social Sciences/Sciences), and university type (key/non-key universities). All my questions have been sufficiently answered. I know I will be invited to review contents of the GSEEE items administered in 2009 through telephone interview. The interview will take 45-60 minutes to complete. An electronic copy of report from the study will be made available to the participants.

I have been notified that participation is voluntary. I may withdraw at any point during the study without affecting my relationship with the investigator. I have been told the steps that will be taken to ensure confidentiality of all information. I have known that there is a minimal risk that my participation will be identified and my comments will have potential influence on my life since the GSEEE test designers (approximately 20) vary each year and the names are highly confidential across a large population of China. However, there is a possibility that my response might be identified by those who already know my opinions. I have been provided with the appropriate contact information in case of questions, concerns, or complaints about participation in this study.

If I have questions, I know that I am free to contact Dr. Liying Cheng, Principal Investigator of the study at 1 (613) 533-6000 ext. 77431 or at liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 1 (613) 533-6081 (chair.GREB@queensu.ca).

Participant’s Name: ______________________
Signature: ________________________________
Date: ________________________________

Please write your e-mail or postal address at the bottom of this sheet if you wish to receive a copy of the results of this study.
Appendix E Content Review Questionnaire

Please examine the Graduate School Entrance English Examination (GSEEE) testing paper you have been given and decide whether each item is likely to advantage/disadvantage test takers in the following groups: 1) male (M) / female (F) 2) Humanities & Social Sciences (HSS) / Sciences (S). Then use the scale below to rate the suitability of those items and write the numbers in the boxes provided. Please explain your choices in the comment area. Thank you for your response.

<table>
<thead>
<tr>
<th>Item</th>
<th>Potential Sources</th>
<th>Scale</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 14</td>
<td>Gender (M / F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text 4 (1-20)</td>
<td>Gender (M / F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text 4 (26-30)</td>
<td>Gender (M / F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 37</td>
<td>Gender (M / F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay writing</td>
<td>Gender (M / F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 30</td>
<td>Academic background (HSS/S)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text 2 (26-30)</td>
<td>Academic background (HSS/S)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Appendix F Item Statistics Output

<table>
<thead>
<tr>
<th>NAME</th>
<th>%right</th>
<th>PCT</th>
<th>Logit</th>
<th>Person</th>
<th>BISERIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 I01</td>
<td>6537.0</td>
<td>48.9</td>
<td>0.03</td>
<td>0.235</td>
<td>0.295</td>
</tr>
<tr>
<td>2 I02</td>
<td>10173.0</td>
<td>76.1</td>
<td>-0.68</td>
<td>0.303</td>
<td>0.416</td>
</tr>
<tr>
<td>3 I03</td>
<td>7109.0</td>
<td>53.2</td>
<td>-0.08</td>
<td>0.342</td>
<td>0.429</td>
</tr>
<tr>
<td>4 I04</td>
<td>6069.0</td>
<td>45.4</td>
<td>0.11</td>
<td>0.194</td>
<td>0.244</td>
</tr>
<tr>
<td>5 I05</td>
<td>7756.0</td>
<td>58.0</td>
<td>-0.19</td>
<td>0.198</td>
<td>0.250</td>
</tr>
<tr>
<td>6 I06</td>
<td>4552.0</td>
<td>34.1</td>
<td>0.39</td>
<td>0.134</td>
<td>0.173</td>
</tr>
<tr>
<td>7 I07</td>
<td>6851.0</td>
<td>51.3</td>
<td>-0.03</td>
<td>0.333</td>
<td>0.417</td>
</tr>
<tr>
<td>8 I08</td>
<td>4478.0</td>
<td>33.5</td>
<td>0.40</td>
<td>0.102</td>
<td>0.133</td>
</tr>
<tr>
<td>9 I09</td>
<td>3135.0</td>
<td>23.5</td>
<td>0.70</td>
<td>0.022</td>
<td>0.030</td>
</tr>
<tr>
<td>10 I10</td>
<td>4804.0</td>
<td>36.0</td>
<td>0.34</td>
<td>0.183</td>
<td>0.235</td>
</tr>
<tr>
<td>11 I11</td>
<td>2063.0</td>
<td>15.4</td>
<td>1.00</td>
<td>0.083</td>
<td>0.126</td>
</tr>
<tr>
<td>12 I12</td>
<td>3844.0</td>
<td>28.8</td>
<td>0.53</td>
<td>-0.037</td>
<td>-0.049</td>
</tr>
<tr>
<td>13 I13</td>
<td>6121.0</td>
<td>45.8</td>
<td>0.10</td>
<td>0.076</td>
<td>0.095</td>
</tr>
<tr>
<td>14 I14</td>
<td>8348.0</td>
<td>62.5</td>
<td>-0.30</td>
<td>0.313</td>
<td>0.400</td>
</tr>
<tr>
<td>15 I15</td>
<td>9063.0</td>
<td>67.8</td>
<td>-0.44</td>
<td>0.228</td>
<td>0.298</td>
</tr>
<tr>
<td>16 I16</td>
<td>3432.0</td>
<td>25.7</td>
<td>0.62</td>
<td>0.099</td>
<td>0.134</td>
</tr>
<tr>
<td>17 I17</td>
<td>5935.0</td>
<td>44.4</td>
<td>0.13</td>
<td>0.110</td>
<td>0.139</td>
</tr>
<tr>
<td>18 I18</td>
<td>5064.0</td>
<td>37.9</td>
<td>0.29</td>
<td>0.166</td>
<td>0.212</td>
</tr>
<tr>
<td>19 I19</td>
<td>5788.0</td>
<td>43.3</td>
<td>0.16</td>
<td>0.174</td>
<td>0.220</td>
</tr>
<tr>
<td>20 I20</td>
<td>9675.0</td>
<td>72.4</td>
<td>-0.57</td>
<td>0.236</td>
<td>0.316</td>
</tr>
<tr>
<td>21 I21</td>
<td>5955.0</td>
<td>44.6</td>
<td>0.13</td>
<td>0.127</td>
<td>0.159</td>
</tr>
<tr>
<td>22 I22</td>
<td>6490.0</td>
<td>48.6</td>
<td>0.03</td>
<td>0.173</td>
<td>0.216</td>
</tr>
<tr>
<td>23 I23</td>
<td>7089.0</td>
<td>53.1</td>
<td>-0.07</td>
<td>0.180</td>
<td>0.226</td>
</tr>
<tr>
<td>24 I24</td>
<td>7547.0</td>
<td>56.5</td>
<td>-0.15</td>
<td>0.285</td>
<td>0.358</td>
</tr>
<tr>
<td>25 I25</td>
<td>6899.0</td>
<td>51.6</td>
<td>-0.04</td>
<td>0.200</td>
<td>0.251</td>
</tr>
<tr>
<td>26 I26</td>
<td>5538.0</td>
<td>41.4</td>
<td>0.20</td>
<td>0.163</td>
<td>0.206</td>
</tr>
<tr>
<td>27 I27</td>
<td>11324.0</td>
<td>84.7</td>
<td>-1.01</td>
<td>0.211</td>
<td>0.321</td>
</tr>
<tr>
<td>28 I28</td>
<td>4478.0</td>
<td>33.5</td>
<td>0.40</td>
<td>0.146</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>29</td>
<td>I29</td>
<td>7992.0</td>
<td>59.8</td>
<td>-0.23</td>
<td>0.243</td>
</tr>
<tr>
<td>30</td>
<td>I30</td>
<td>9536.0</td>
<td>71.4</td>
<td>-0.54</td>
<td>0.113</td>
</tr>
<tr>
<td>31</td>
<td>I31</td>
<td>9060.0</td>
<td>67.8</td>
<td>-0.44</td>
<td>0.345</td>
</tr>
<tr>
<td>32</td>
<td>I32</td>
<td>11394.0</td>
<td>85.3</td>
<td>-1.03</td>
<td>0.270</td>
</tr>
<tr>
<td>33</td>
<td>I33</td>
<td>6700.0</td>
<td>50.1</td>
<td>0.00</td>
<td>0.036</td>
</tr>
<tr>
<td>34</td>
<td>I34</td>
<td>6955.0</td>
<td>52.1</td>
<td>-0.05</td>
<td>0.123</td>
</tr>
<tr>
<td>35</td>
<td>I35</td>
<td>8190.0</td>
<td>61.3</td>
<td>-0.27</td>
<td>0.162</td>
</tr>
<tr>
<td>36</td>
<td>I36</td>
<td>3780.0</td>
<td>28.3</td>
<td>0.55</td>
<td>0.105</td>
</tr>
<tr>
<td>37</td>
<td>I37</td>
<td>5600.0</td>
<td>41.9</td>
<td>0.19</td>
<td>0.169</td>
</tr>
<tr>
<td>38</td>
<td>I38</td>
<td>10191.0</td>
<td>76.3</td>
<td>-0.69</td>
<td>0.197</td>
</tr>
<tr>
<td>39</td>
<td>I39</td>
<td>8224.0</td>
<td>61.5</td>
<td>-0.28</td>
<td>0.241</td>
</tr>
<tr>
<td>40</td>
<td>I40</td>
<td>7247.0</td>
<td>54.2</td>
<td>-0.10</td>
<td>0.156</td>
</tr>
<tr>
<td>41</td>
<td>I41</td>
<td>8612.0</td>
<td>64.5</td>
<td>-0.35</td>
<td>0.312</td>
</tr>
<tr>
<td>42</td>
<td>I42</td>
<td>9121.0</td>
<td>68.3</td>
<td>-0.45</td>
<td>0.262</td>
</tr>
<tr>
<td>43</td>
<td>I43</td>
<td>1171.0</td>
<td>8.8</td>
<td>1.38</td>
<td>-0.065</td>
</tr>
<tr>
<td>44</td>
<td>I44</td>
<td>2714.0</td>
<td>20.3</td>
<td>0.80</td>
<td>0.019</td>
</tr>
<tr>
<td>45</td>
<td>I45</td>
<td>5404.0</td>
<td>40.4</td>
<td>0.23</td>
<td>0.186</td>
</tr>
</tbody>
</table>
Appendix G Letter of Information for Test Takers

“Test stakeholder perceptions on fairness of a large-scale high-stakes language test”

I, Xiaomei Song, am a doctoral student in the Faculty of Education, Queen’s University, Canada. I will be conducting a research project on “Test stakeholder perceptions on fairness of a large-scale high-stakes language test”. This project will investigate test stakeholder perceptions on fairness of the Chinese Graduate School English Entrance Examination (GSEEE). This study has been granted clearance by the General Research Ethics Board for compliance with the Tri-Council Policy Statements: Ethical Conduct of Research Involving Humans, and Queen’s policies.

You are invited to participate in the study since you plan to take the GSEEE of the 2012 administration. I will invite a group of test takers like you (about 4 to 6) to conduct focus group interviewing. I will conduct a three-part interview after you take the GSEEE. The first part is about your background information, the second part about your knowledge and experience with the GSEEE, and the third part about your perceptions on fairness of the GSEEE. The focus group interview will last 60-90 minutes. After you are notified of the GSEEE test score by the National Education Examinations Authority (NEEA), I will do a follow-up with you by phone or email regarding your perceptions on scoring and ask you to provide other comments if you like. The follow-up will take less than 5 minutes. The interview and the follow-up will be recorded and transcribed.

Your involvement in this research is entirely voluntary and there are no known or anticipated risks to participate in this research. You can withdraw from the research at any time without reasons and without any consequences to you. You may decline answering any interview questions you find objectionable or uncomfortable. The transcribed interview data will be sent to you for verification. Prior to the focus group interview, I will ask the focus group participants to respect the confidentiality of their fellow participants in the focus group. They will be told that they are not to discuss the content of the focus group interview outside of the focus group meeting. During the focus group interview, each individual will be speaking about his or her situation in front of other members of the focus group. All information you provide will be kept confidential to the extent possible. My research assistant who takes notes and deals with the recorder will sign a confidentiality agreement. A pseudonym will be assigned to you for the identification purpose and to protect confidentiality. Your personal information will not be attached to any form of the research data. The data will be kept in a locked cabinet/computer in the researcher’s office. Only the researcher, the committee, and supervisor will have access to data. In accordance with Queen’s policy, research data will be retained for five years. Then, it will be destroyed.

This study may result in publications of various types, such as conference presentations and journal articles. The data of the study may also be used as secondary data analysis. Your name will not be attached to any form of the data that you provide, nor will it appear in any publication created as a result of this study. An electronic copy of a report
from this study will be made available to you upon your request. Any questions about study participation may be directed to the researcher, Xiaomei Song, at 001-519-9678470 (0xs@queensu.ca), or you can contact my supervisor, Dr. Liying Cheng, at 001-613-533-6000 ext. 77431 (living.cheng@queensu.ca). Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 001-613-533-6081 (chair.GREB@queensu.ca).

Thank you for your interest in this study, Xiaomei Song, PhD Candidate, Faculty of Education, Queen’s University, Canada
I, Xiaomei Song, am a doctoral student in the Faculty of Education, Queen’s University, Canada. I will be conducting a research project on “Test stakeholder perceptions on fairness of a large-scale high-stakes language test”. This project will investigate test stakeholder perceptions on fairness of the Chinese Graduate School English Entrance Examination (GSEEEE). This study has been granted clearance by the General Research Ethics Board for compliance with the Tri-Council Policy Statements: Ethical Conduct of Research Involving Humans, and Queen’s policies.

You are invited to participate in this research since you are the key administrator of your graduate program and have been involved in making decisions based on the GSEEEE test scores. I will conduct a three-part interview with you. The first part is about your background information, the second part about your knowledge and experience with the GSEEEE, and the third part about your perceptions on fairness of the GSEEEE. The survey will last 45-60 minutes. The interview will be recorded and transcribed.

Your involvement in this research is entirely voluntary and there are no known or anticipated risks to participate in this research. You can withdraw from the research at any time without reasons and without any consequences to you. You may decline answering any interview questions you find objectionable or uncomfortable. The transcribed interview data will be sent to you for verification. A pseudonym will be assigned to you for the identification purpose and to protect confidentiality. All information you provide will be kept confidential to the extent possible. Your personal information will not be attached to any form of the research data. The data collected will be kept in a cabinet/computer in the researcher’s office. Only the researcher, the committee, and supervisor will have access to data. In accordance with Queen’s policy, research data will be retained for five years. Then, it will be destroyed.

This study may result in publications of various types, such as conference presentations and journal articles. The data of the study may also be used as secondary data analysis. Your name will not be attached to any form of the data that you provide, nor will it appear in any publication created as a result of this study. An electronic copy of a report from this study will be made available to you upon your request. Any questions about study participation may be directed to the researcher, Xiaomei Song, at 001-519-9678470 (0xs@queensu.ca), or you can contact my supervisor, Dr. Liying Cheng, at 001-613-533-6000 ext. 77431 (liying.cheng@queensu.ca). Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 001-613-533-6081 (chair.GREB@queensu.ca).

Thank you for your interest in this study, Xiaomei Song, PhD Candidate, Faculty of Education, Queen’s university, Canada
“Test stakeholder perceptions on fairness of a large-scale high-stakes language test”

I, Xiaomei Song, am a doctoral student in the Faculty of Education, Queen’s University, Canada. I will be conducting a research project on “Test stakeholder perceptions on fairness of a large-scale high-stakes language test”. This project will investigate test stakeholder perceptions on fairness of the Chinese Graduate School English Entrance Examination (GSEEE). This study has been granted clearance by the General Research Ethics Board for compliance with the Tri-Council Policy Statements: Ethical Conduct of Research Involving Humans, and Queen’s policies.

You are invited to participate in the research since you have taught English courses at Master’s level. I will conduct a three-part interview with you. The first part is about your background information, the second part about your knowledge and experience with the GSEEE, and the third part about your interpretations and perceptions on fairness of the GSEEE. The survey will last 45-60 minutes. The interview data will be recorded and transcribed.

Your involvement in this research is entirely voluntary and there are no known or anticipated risks to participate in this research. You can withdraw from the research at any time without reasons and without any consequences to you. You may decline answering any interview questions you find objectionable or uncomfortable. The transcribed interview data will be sent to you for verification. A pseudonym will be assigned to you for the identification purpose and to protect confidentiality. All information you provide will be kept confidential to the extent possible. Your personal information will not be attached to any form of the research data. The data collected will be kept in a cabinet/computer in the researcher’s office. Only the researcher, the committee, and supervisor will have access to this data. In accordance with Queen’s policy, research data will be retained for five years. Then, it will be destroyed.

This study may result in publications of various types, such as conference presentations and journal articles. The data of the study may also be used as secondary data analysis. Your name will not be attached to any form of the data that you provide, nor will it appear in any publication created as a result of this study. An electronic copy of a report from this study will be made available to you upon your request. Any questions about study participation may be directed to the researcher, Xiaomei Song, at 001-519-9678470 (0xs@queensu.ca), or you can contact my supervisor, Dr. Liying Cheng, at 001-613-533-6000 ext. 77431 (liying.cheng@queensu.ca). Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 001-613-533-6081 (chair.GREB@queensu.ca).

Thank you for your interest in this study, Xiaomei Song, PhD Candidate, Faculty of Education, Queen’s University, Canada
Appendix J Consent Form for Test Takers

I have read, understood, and retained a copy of the Letter of Information concerning “Test stakeholder perceptions on fairness of a large-scale high-stakes language test”. All my questions regarding the research project have been sufficiently answered. I am aware of the purposes and data collection procedures in this research. I understand the purpose of the research is to investigate test stakeholder perceptions on fairness of the Chinese Graduate School English Entrance Examination (GSEEE). I understand that I will be interviewed with a group of test takers like me for about 60-90 minutes after I take the GSEEE of the 2012 administration. Then, after I have been notified of my GSEEE test score by the National Education Examinations Authority (NEEA), the researcher will do a follow-up with me by phone or email regarding my perceptions on scoring and ask me to provide other comments if I like. The follow up will take less than 5 minutes. The interview data and the follow up will be recorded and transcribed.

I have been notified that my participation in this study is entirely voluntary and that I may withdraw at any point during the study. Should I choose to withdraw I may request removal of all or part of my interview data, without having any consequences to me. I understand that I can choose not to answer any interviewed questions that I find objectionable or uncomfortable. A pseudonym will be assigned to me for the identification purpose and to protect my confidentiality. I have been told the steps that will be taken to keep confidentiality of all information to the extent possible. I understand that I will be speaking about my personal situation in front of other members of the focus group and I have been advised not to disclose information that I hear in the focus groups outside the focus group meeting. I have also been informed that the researcher will ask focus group participants to respect the confidentiality of their fellow participants in the focus group. I have been provided with the appropriate contact information in case of questions, concerns, or complaints about participation in this study.

Any questions about study participation may be directed to Xiaomei Song, at 001-519-9678470 (0xs@queensu.ca), or her supervisor Dr. Liying Cheng at 001-613-533-6000, ext: 77431 (living.cheng@queensu.ca). Any ethical concerns about the research may be directed to the Chair of the General Research Ethics Board, telephone at 001-613-533-6081 (chair.GREB@queensu.ca).

Please sign one copy of this Consent Form and return to Xiaomei Song. Retain the second copy for your records.

Participant’s Name: _____________________________________________________

Signature: _____________________________________________________________

Date: ________________________________________________________________
E-mail address for receiving interview transcription and copy of findings:
______________________________________________________
Appendix K Consent Form for Program Administrators

I have read, understood, and retained a copy of the Letter of Information concerning “Test stakeholder perceptions on fairness of a large-scale high-stakes language test”. All my questions regarding the study have been sufficiently answered. I am aware of the purposes and data collection procedures in this research. I understand the purpose of the research is to examine test stakeholder perceptions on fairness of the Chinese Graduate School Entrance English Exam (GSEE). I understand that I will complete a one-on-one interview for about 45-60 minutes. The interview data will be recorded and transcribed.

I have been notified that my participation in this study is entirely voluntary and that I may withdraw at any point during the study. Should I choose to withdraw I may request removal of all or part of my interview data, without having any consequences to me. I understand that I can choose not to answer any interviewed questions that I find objectionable or uncomfortable. A pseudonym will be assigned to me for the identification purpose and to protect my confidentiality. I have been told the steps that will be taken to keep confidentiality of all information to the extent possible. I have been provided with the appropriate contact information in case of questions, concerns, or complaints about participation in this study.

Any questions about study participation may be directed to Xiaomei Song, at 001-519-9678470 (0xs@queensu.ca), or her supervisor Dr. Liying Cheng at 001-613-533-6000, ext: 77431 (liying.cheng@queensu.ca). Any ethical concerns about the research may be directed to the Chair of the General Research Ethics Board, telephone at 001-613-533-6081 (chair.GREB@queensu.ca).

Please sign one copy of this Consent Form and return to Xiaomei Song. Retain the second copy for your records.

Participant’s Name: _____________________________________________________

Signature: __________________________________________________________________

Date: _____________________________________________________________________

E-mail address for receiving interview transcription and copy of findings:
_________________________________________________________________________
Appendix L Consent Form for English Teachers

I have read, understood, and retained a copy of the Letter of Information concerning “Test stakeholder perceptions on fairness of a large-scale high-stakes language test”. All my questions regarding the study have been sufficiently answered. I am aware of the purposes and data collection procedures in this research. I understand the purpose of the research is to examine test stakeholder perceptions on fairness of the Chinese Graduate School Entrance English Exam (GSEE). I understand that I will complete a one-on-one interview for about 45-60 minutes. The interview data will be recorded and transcribed.

I have been notified that my participation in this study is entirely voluntary and that I may withdraw at any point during the study. Should I choose to withdraw I may request removal of all or part of my interview data, without having any consequences to me. I understand that I can choose not to answer any interviewed questions that I find objectionable or uncomfortable. A pseudonym will be assigned to me for the identification purpose and to protect my confidentiality. I have been told the steps that will be taken to keep confidentiality of all information to the extent possible. I have been provided with the appropriate contact information in case of questions, concerns, or complaints about participation in this study.

Any questions about study participation may be directed to Xiaomei Song, at 001-519-9678470 (0xs@queensu.ca), or her supervisor Dr. Liying Cheng at 001-613-533-6000, ext: 77431 (liying.cheng@queensu.ca). Any ethical concerns about the research may be directed to the Chair of the General Research Ethics Board, telephone at 001-613-533-6081 (chair.GREB@queensu.ca).

**Please sign one copy of this Consent Form and return to Xiaomei Song. Retain the second copy for your records.**

Participant’s Name: ___________________________________________________

Signature: ______________________________________________________________________________

Date: ______________________________________________________________________________________

E-mail address for receiving interview transcription and copy of findings:

___________________________________________________________________________________________
Appendix M Recruitment Poster for Focus Group with Test takers

Research Participants Invited

I, Xiaomei Song, am a doctoral student in the Faculty of Education, Queen’s University, Canada. I am conducting a research project on “Test stakeholders’ perceptions of the fairness of a large-scale high-stakes language test”. This project investigates test stakeholder perceptions on fairness of the Chinese Graduate School English Entrance Examination (GSEEE).

If you will take the GSEEE administered in 2012, I invite you to participate in the study. You will be asked about your perceptions on fairness of the GSEEE through focus group interviewing after you take the test. The focus group interview will last about 60 to 90 minutes. After you have been notified of the GSEEE test score by the National Education Examinations Authority (NEEA), I will do a follow-up with you by phone or email regarding your perceptions on scoring and ask you to provide other comments if you like. The follow-up will take less than 5 minutes.

If you are interested in participating the study, please contact Xiaomei Song at 001-519-9678470 or by email 0xs@queensu.ca

Thank you very much for your interest.
Appendix N Interview Protocol with Program Administrators

Section One
1. Can you introduce yourself (e.g., name, gender, age)?
2. How long have you been involved in the graduate schools or programs? In what role(s)?
3. What specific role(s) do you play in the admission of Master’s students?

Section Two
4. Can you describe your experience with the GSEEE? What are the specific requirements and steps in terms of the selection process in your program?
5. Can you tell me what you know about the GSEEE, for example, format, purpose, administration, scoring, test use, etc?

Section Three
6. Based on your knowledge and experience, how do you think the GSEEE to predict different groups of students’ readiness in Master’s learning?
7. Based on your knowledge and experience, how does your program use the GSEEE test scores? How do you think about the fairness of the decisions for test takers based on their GSEEE test score and other scores?
8. What do you think of the fairness of the GSEEE on the whole? And why?
9. Do you have any other comments?
Appendix O Interview Protocol with English Teachers

**Section One**
1. Can you introduce yourself (e.g., name, gender, age)?
2. What teaching experience do you have? How long have you been involved in teaching English classes for Master’s students?
3. What specific role(s) do you play in the use of the GSEEE test scores?

**Section Two**
4. Can you describe your experience with the GSEEE?
5. Can you tell me what you know about the GSEEE, for example, format, purpose, administration, scoring, test use, etc.?

**Section Three**
6. Based on your knowledge and experience, how do you think the GSEEE to predict different groups of students’ readiness in Master’s learning?
7. Based on your knowledge and experience, how does your school use the GSEEE test scores? How do you think about the fairness of the decisions for test takers based on their GSEEE test score and other scores?
8. What do you think of the fairness of the GSEEE on the whole? And why?
9. Do you have any other comments?
Appendix P Interview Protocol for Focus Group with Test takers

Section One
1. Can you introduce yourself (e.g., name, gender, age)?
2. Which school and program have you applied for this year?
3. What do you think about your overall English proficiency (relatively low, medium, relatively high)? In reading? Writing? Listening? And Speaking?

Section Two
4. Can you describe your experience with the GSEEE? Is this your first time to take the GSEEE?
5. Can you tell me what you know about the GSEEE, for example, format, purpose, administration, scoring, test use, etc.? Can you describe your preparation and test registration experience?

Section Three
6. Thinking back the 2012 administration you have taken, how do you think that the test and test items, say, in terms of content, format, and tested skills, provided test takers like you (e.g., female, Sciences background) with fair opportunities to demonstrate your English proficiency? Can you give me an example?
7. Thinking back the 2012 administration you have taken, how do you think that the GSEEE test administration provided you with fair opportunities to demonstrate your English proficiency? Can you give me an example?
8. You have recently received the official report about your GSEEE score. How do you think about your score on the GSEEE? Do you feel your performance on the GSEEE is similar to what you scored for yourself?
9. How do you think about the fairness of the decisions for test takers based on their GSEEE test score and other scores? Do you think the GSEEE to predict different groups of students’ readiness in Master’s learning?
10. What do you think of the fairness of the GSEEE on the whole? And why?
11. Do you have any other comments?