CONTEXT, CONSTRUCT, AND CONSEQUENCES:
WASHBACK OF THE COLLEGE ENGLISH TEST IN CHINA

by

Youyi Sun

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
(January, 2016)

Copyright ©Youyi Sun, 2016
Abstract

The College English Test (CET) is a large-scale language testing program implemented for the purpose of enhancing English as a foreign language (EFL) in the Chinese tertiary education context. Its results have been used for different purposes in varied contexts. The multiple uses have generated growing concerns and debate over consequences and validity. However, in the field of language assessment, there are few empirical studies that investigate test consequences within a coherent validation framework. Without such studies, it’s hard to know how and why the CET program influences teaching and learning. To address this research gap and research problem, this multi-phase, multi-method study investigated the intended and unintended consequences of the CET, focusing on its washback on students’ learning and learning outcomes. Drawing on the argument-based validation approach, I linked evidence about the CET consequences to construct validity. Participants of this study were multiple stakeholders including the test developer, test users and students. Interview, questionnaire, and test score data were collected. The interview data were analyzed thematically. Structural equation modeling (SEM) was used to analyze the questionnaire and the test score data. Findings of this study highlight the complexity of the consequences of the CET in the Chinese educational and societal contexts. This complexity was attributable to both properties of the test and other contextual forces, particularly uses of the test in the educational and societal contexts. These findings indicate that while the CET provides useful information about students’ achievement of the curriculum, the decision to use the testing program to enhance EFL education is not supported by evidence regarding its consequences. This study sheds light on understanding why and in what ways the CET washback exists in the Chinese educational context. Findings of this study have significant implications for future washback research and for the CET implementation as well as EFL teaching and learning.
Acknowledgements

For me, undertaking a PhD is challenging physically, emotionally, intellectually and financially. Thanks to the support and encouragement I received from my advisory committee, colleagues, friends, and family, this process has become an exciting and rewarding journey.

First and foremost, I would like to express my sincere gratitude to my supervisor and mentor Dr. Liying Cheng. Seven years ago, an email message from Dr. Cheng changed my life forever. Ever since then, her support has never failed. Her guidance has kept me on the right track and at the same time her encouragement enables me to explore the complexity of consequences, washback and validity from different perspectives. Our conversations about life and academia have shaped my life and career. Our collaborations in co-authored papers, conference presentations, grant applications, and teaching are the highlights of my career in the past years. I hope these conversations and collaborations continue.

When it came time to select my dissertation committee, I knew Dr. Lyn Shulha and Dr. Don Klinger could give me the strongest support, but not sure whether they would accept my request considering their extremely busy schedule and workload. To my great excitement, both of them agreed immediately. The Program Evaluation course I took from Dr. Lyn Shulha inspired me to conduct research into consequences and washback drawing on insights from program evaluation. Since I started working on my doctoral dissertation proposal, Dr. Shulha has continuously brought clarity and perspective to my work through insightful and detailed feedback, even after her retirement. I owe my current job position to the expertise and skills I gained in the courses offered by Dr. Don Klinger and to his recommendation. Dr. Klinger’s support to my dissertation is exceptional. His suggestion of using Kane’s model was a great help.
in terms of methodological framework. Dr. Klinger’s critical questions did not scare me. Rather, they challenged me to think deeper about the research and make me stronger academically.

I would like to acknowledge the contributions of the examiners and the Dean’s Delegate at my oral examination: Dr. Alister Cumming, Dr. Lee Fabrigar, and Dr. Richard Reeve. The challenging, thought-provoking questions they posed broadened my thinking about this study. I am grateful to Dr. Cumming for his encouragement and careful and detailed feedback.

My studies at the Faculty of Education have been enriched by numerous faculty and staff members in both formal and informal ways. Thanks to all of you for your support along the way. Special thanks go to the Assessment and Evaluation Group, particularly Dr. Christine Doe and Dr. Christopher Deluca. Thank you for your insightful comments on my papers, proposals and presentations, the rides you offered me, and your generous support to my job applications.

My special thanks go to Dr. Janna Fox, who introduced me to the language assessment world when I was studying at Carleton University, Ottawa. Without her inspiration, encouragement, and continuous recommendation, my PhD studies at Queen’s University would not be possible. My thanks also go to Dr. Lynne Young at Carleton University, who always encouraged me to pursue my PhD.

My heartfelt thanks to my family: my wife Shuzhen and my son Haoyu. I know I owe you both too much over the past years. Thank you for your love and understanding when I was not able to afford more time with you. Thank you, Shuzhen, for taking care of most of the housework and for helping me with the data entry and editing my papers. Thank you, Haoyu, for giving me support in statistics. My thanks also go to my parents-in-laws for their support and care. A special thank you to Helena Yu, who gave me most generous support and made me feel
at home while studying in Kingston. I also want to thank Hang Li and Shumei He for their support to me and my family in Kingston.

I want to thank my colleagues for their great support. Special thanks go to my former colleague Dr. He Ying, vice-president of Shanghai Finance University and my current colleague Ken Clark, coordinator of the Assessment Unit of Instruction, Curriculum and Assessment Branch, Manitoba Education and Advanced Learning, for their great support to my work so that I was able to have more time and energy to study in the PhD program and to work on this dissertation.

Numerous friends have provided generous help in my doctoral studies. Special thanks go to Chen Ping, for not only helping me contact participants for the study but also preparing very delicious food during the data collection. I also want to thank Jianhua, Guilin, Sun Wen, Yonggang, and Yu Feng, my best friends in Jining University for their help in the data collection process. I am grateful to all the anonymous participants in this study. Thank you all for your participation.

I am grateful for the financial support I received to support my research and funding for my PhD program including but not limited to:

- 2011-2012 Ontario Graduate Scholarships
- 2012 TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment: from Educational Testing Service
- 2010-2011 Graduate Dean's Doctoral Field Travel Grant from Queen’s University

My sincere thanks to all.
# Table of Contents

Abstract ................................................................................................................................. ii
Acknowledgements .............................................................................................................. iii
List of Figures ....................................................................................................................... ix
List of Tables ....................................................................................................................... x
List of Abbreviations .......................................................................................................... xi

Chapter 1 Introduction ........................................................................................................ 1
  1.1 Purposes ....................................................................................................................... 1
  1.2 Conceptualizations of Consequences, Impact, and Washback ................................. 2
  1.3 Context of This Study ............................................................................................... 3
      1.3.1 The College English Test ............................................................................... 4
      1.3.2 Context of the CET Implementation ............................................................. 6
  1.4 Rationale of the Study ............................................................................................. 11
  1.5 Significance ............................................................................................................... 12
  1.6 Dissertation Structure ............................................................................................. 14

Chapter 2 Literature Review ............................................................................................ 17
  2.1 Test Validation and Consequences ......................................................................... 17
  2.2 Application of the Argument-based Validation Approach in This Study ............ 24
  2.3 Research into Consequences of Large-Scale Assessment ....................................... 27
      2.3.1 Intended and Unintended Direct Consequences ........................................... 27
      2.3.2 Intended and Unintended Indirect Consequences in the Educational System ... 33
  2.4 Impact and Washback Research in Language Assessment ..................................... 36
      2.4.1 Impact Research in Language Assessment .................................................. 37
      2.4.2 Washback Research in Language Assessment .............................................. 42
  2.5 Chapter Summary .................................................................................................... 50

Chapter 3 Methodology .................................................................................................... 53
  3.1 Multi-phase, Multi-method Research Design of the Study ..................................... 54
  3.2 Researcher’s Role in This Study ............................................................................. 57
  3.3 Operational Definitions of Terms in Research Questions ....................................... 58
  3.4 Methods .................................................................................................................... 59
      3.4.1 Participants ..................................................................................................... 60
      3.4.2 Instruments .................................................................................................... 64
      3.4.3 Data Collection ............................................................................................. 67
Chapter 4 Intended Consequences and Actual Uses of the CET-4: Test Developer’s and Test Users’ Perspectives

4.1 Intended Consequences of the CET-4

4.1.1 To Provide Useful Information about Students’ English Proficiency

4.1.2 To Promote College English Education in China

4.2 Procedures to Address Consequences of the CET-4

4.2.1 Procedures Are Taken to Ensure Quality of the CET-4

4.2.2 Procedures Are Taken to Address Test Consequences

4.3 Test Uses in Educational and Societal Contexts

4.3.1 Test Uses in the Educational Context

4.3.2 Test Uses in the Societal Contexts

4.4 Value Implications Underlying Uses of the CET-4 Results

4.4.1 Values Associated with the CET-4 Construct

4.4.2 Values Associated with the CET-4

4.5 Chapter Summary

Chapter 5 Washback of the CET-4 on Learning and Learning Outcomes

5.1 Perceptions of CET-4 Demands and Uses, Test-taking Expectations, and Test Preparation Practices

5.1.1 Perceptions of CET-4 Demands

5.1.2 Perceptions of CET-4 Score Uses

5.1.3 Test-taking Expectations

5.1.4 Test Preparation Practices

5.2 Test Performance

5.3 Relationships Among Perceptions of CET-4, Test-taking Expectations, Test Preparation Practices, and Test Performance

5.3.1 The CET-4 Washback Model

5.3.2 Evaluation of the Global Fit of the CET-4 Washback Model

5.3.3 Parameter Estimation for the CET-4 Washback Model

Chapter 6 Discussion and Conclusions

6.1 Test Developer’s Intended Consequences of the CET-4

6.1.1 Promoting College English Education in China

6.1.2 Informing College English Education

6.2 Test Developer’s Efforts to Achieve Intended Consequences
6.2.1 Efforts in Support of the Decision Rule ......................................................... 153
6.2.2 Efforts in Support of the Utilization Inference ........................................... 156
6.3 Uses of the CET-4 in the Educational and Societal Contexts ....................... 163
   6.3.1 Uses of the CET-4 in the Educational Context ....................................... 163
   6.3.2 Uses of the CET-4 in the Business and the Government Contexts .......... 165
   6.3.3 Value Implications Underlying Actual Uses of the CET-4 ....................... 167
6.4 Washback of the CET-4 on Learning ............................................................. 170
   6.4.1 Intended and Unintended Washback of the CET-4 on Learning ............... 170
   6.4.2 Factors Contributing to the Washback of the CET-4 .............................. 174
6.5 Conclusions ................................................................................................. 176
6.6 Limitations and Directions for Future Research ........................................... 178
6.7 Contributions and Implications for Test Development .................................. 180

References ........................................................................................................ 183

Appendix A. Ethics Clearance Approval Letter .................................................... 211
Appendix B. Consent Form to Test Developer Participant ..................................... 212
Appendix C. Consent Form to Test User Participants .......................................... 213
Appendix D. Consent Form to Student Participants ............................................. 214
Appendix E. Interview Guideline with the CET Developer Participant ................. 215
Appendix F. Interview Guideline with Test User Participants ............................. 216
Appendix G. First Page of CET-4 sample Test Paper ......................................... 217
Appendix H. Questionnaire for Student Participants .......................................... 218
Appendix I. Excerpt from the CET Website about Score Interpretation ............ 225
Appendix J. Covariance Matrix for the Construct Indicators in the CET-4 Washback Model ... 226
List of Figures

Figure 1-1. Structure of the dissertation ................................................................. 14
Figure 2-1. Structure of assessment use argument .................................................. 22
Figure 2-2. An IUA Framework for the CET-4......................................................... 24
Figure 2-3. Delineating the responsibilities of test developers and test users .............. 30
Figure 2-4. Mechanisms affecting language policy .................................................. 41
Figure 2-5. A basic model of washback .................................................................... 44
Figure 2-6. Model of washback incorporating direction and intensity .......................... 46
Figure 2-7 A Washback Model of Student Learning ................................................ 47
Figure 3-1. Multi-phase, multi-method research design of the study ......................... 54
Figure 3-2. A thematic analysis system .................................................................... 72
Figure 3-3. Demonstration of three coding levels ....................................................... 73
Figure 3-4. Structural regression model of the CET-4 washback ............................... 76
Figure 5-1. Proposed CET-4 Washback Model ....................................................... 119
Figure 5-2. CET-4 washback model generated by LISREL ....................................... 141
Table 1.1  Structure of the CET-4 Test Paper and Administration Time ............................................... 6
Table 2.1  Unitary Validity Model ........................................................................................................... 20
Table 3.1  Profile of Test User Participants ............................................................................................. 62
Table 3.2  Scales and Sub-Scales Measured by the Questionnaire .......................................................... 66
Table 4.1  Intended Consequences of the CET-4 .................................................................................... 79
Table 4.2  Matrix of Findings on Procedures to Achieve Intended Consequences ................................. 86
Table 4.3  Basic Level Requirements in the College English Curriculum .............................................. 90
Table 4.4  Matrix of Findings on Uses of CET-4 Results ....................................................................... 97
Table 4.5  Matrix of Findings on Value Implications Underlying Test Uses ......................................... 109
Table 5.1  Descriptive Statistics of Perceptions of CET-4 Demands ....................................................... 121
Table 5.2  Rotated Factor Matrix of Perceptions of CET-4 Demands ................................................... 124
Table 5.3  Descriptive Statistics of Factors Concerning Perceptions of CET-4 Task Demands .................. 125
Table 5.4  Descriptive Statistics of Perceptions of CET-4 Uses ................................................................. 126
Table 5.5  Rotated Factor Matrix of Perceptions of CET-4 Uses ............................................................... 127
Table 5.6  Descriptive Statistics of Factors Concerning Perceptions of CET-4 Uses ............................... 128
Table 5.7  Descriptive Statistics of Students’ Expectations of Taking CET-4 ............................................ 129
Table 5.8  Rotated Factor Matrix of Test Taking Expectations ................................................................. 130
Table 5.9  Descriptive Statistics of Factors on Students’ Expectations of Taking CET-4 ...................... 131
Table 5.10 Descriptive Statistics of CET-4 Test Preparation ................................................................. 132
Table 5.11 Rotated Factor Matrix of CET-4 Test Preparation ............................................................... 134
Table 5.12 Descriptive Statistics of Factors on CET-4 Test Preparation ............................................... 135
Table 5.13 Descriptive Statistics of the CET-4 Section-level Test Score Data ..................................... 136
Table 5.14 Rotated Factor Matrix of Test Score Data ............................................................................. 137
Table 5.15 Descriptive Statistics for Learning Outcomes .......................................................................... 138
Table 5.16 Latent Variables and Indicators in the CET-4 Washback Model ............................................ 139
Table 5.17 Correlation Matrix for Construct Indicators in the CET-4 Washback Model ....................... 140
Table 5.18 Global Fit Indices for the CET-4 Washback Model .............................................................. 143
Table 5.19 Parameter and Standard Error Estimates for the CET-4 Washback Model ......................... 144
Table 6.1  Research Questions and Findings in Relation to the Inferences in the IUA ............................. 147
List of Abbreviations

ACU: achievement uses
AGFI: adjusted goodness-of-fit index
AUA: assessment use argument
BAK: background knowledge
BEC: Business English Certificates
CAELD: the Canadian Academic English Language Diagnostic
CECR: College English Curriculum Requirements
CERP: College English Reform Program
CET: College English Test
CFI: comparative fit index
CMoE: Chinese Ministry of Education
EAP: English for academic purposes
EFL: English as a foreign language
EGP: English for general purposes
ELLs: English language learners
EQAO: Education Quality and Accountability Office
Expec: test-taking expectations
FWT: fast reading, writing, and translation
GFE: goal-free evaluation
HEQRP: Higher Education Quality and Reform Project
IELTS: International English Language Testing System
INU: instrumental uses
IS: integrated skills
IUA: interpretation/use argument
KMO: Kaiser-Meyer-Olkin
LAN: Local Area Network
LC: listening comprehension (a section of the CET-4)
LIC: listening comprehension (a sub-construct measured by the CET-4)
LIS: listening skills
LKS: language knowledge and skills
LSD: long-term skill development
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQs</td>
<td>multiple choice questions</td>
</tr>
<tr>
<td>NCEE</td>
<td>National College Entrance Examinations</td>
</tr>
<tr>
<td>NCETC</td>
<td>National College English Testing Committee</td>
</tr>
<tr>
<td>NMET</td>
<td>National Matriculation English Test</td>
</tr>
<tr>
<td>OSSLT</td>
<td>Ontario Secondary School Literacy Test</td>
</tr>
<tr>
<td>PAF</td>
<td>Principal axis factoring</td>
</tr>
<tr>
<td>Pdemand</td>
<td>perceptions of test demands</td>
</tr>
<tr>
<td>Perform</td>
<td>test performance</td>
</tr>
<tr>
<td>Prep</td>
<td>test preparation practices</td>
</tr>
<tr>
<td>Puses</td>
<td>perceptions of test uses</td>
</tr>
<tr>
<td>RAC</td>
<td>rehearsal and cramming</td>
</tr>
<tr>
<td>RC</td>
<td>reading comprehension (a section of the CET-4)</td>
</tr>
<tr>
<td>REC</td>
<td>reading comprehension (a sub-construct measured by the CET-4)</td>
</tr>
<tr>
<td>RES</td>
<td>reading skills</td>
</tr>
<tr>
<td>RMSEA</td>
<td>root-mean-square error of approximation</td>
</tr>
<tr>
<td>SAS</td>
<td>social-affective strategies</td>
</tr>
<tr>
<td>SEE</td>
<td>self-efficacy</td>
</tr>
<tr>
<td>SEM</td>
<td>structural equation modeling</td>
</tr>
<tr>
<td>SET</td>
<td>Spoken English Test</td>
</tr>
<tr>
<td>SIA</td>
<td>social impact analysis</td>
</tr>
<tr>
<td>SILT</td>
<td>Studies in Language Testing</td>
</tr>
<tr>
<td>SRL</td>
<td>self-regulated learning</td>
</tr>
<tr>
<td>SRMR</td>
<td>standardized root mean square residual</td>
</tr>
<tr>
<td>TAD</td>
<td>task difficulty</td>
</tr>
<tr>
<td>TAV</td>
<td>task values</td>
</tr>
<tr>
<td>TLU</td>
<td>target language use</td>
</tr>
<tr>
<td>TPM</td>
<td>test preparation management</td>
</tr>
<tr>
<td>wpm</td>
<td>words per minute</td>
</tr>
<tr>
<td>WT</td>
<td>writing and translation</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

1.1 Purposes

The CET is an English as a foreign language (EFL) test that is taken by the largest number of students in the world, with an annual test taker population of over 18 million (Jin, 2011). Its results are extensively used in different contexts such as education, business, and government to make high-stakes decisions about students. Consequently, the CET directly or indirectly influences students’ study and life chances. Low scores can prevent students from obtaining academic degrees, entering into professions they are prepared for, or even obtaining residence permits\(^1\) in some major cities in China. Largely because of controversies surrounding its impact, the CET is “probably the most debated test in the language testing field and among academics in China” (Cheng, 2008b, p. 33).

Drawing on Kane’s (2006) argument-based validation approach, this study evaluates the decision of using the CET program as a policy instrument for educational reform, focusing on the consequences of the CET-4 (see next section) in the Chinese educational and societal contexts, particularly its washback (i.e., its influences on students’ learning and learning outcomes). The purpose of this evaluation is to address the research problem why the CET-4 washback exists in the Chinese tertiary education context.

Test validation is in nature a process of evaluation (Cronbach, 1988, 1989). Evaluation of large-scale assessment programs that are implemented to enhance

\(^1\) A residence permit is called *Hukou* in Chinese. *Hukou* in some major cities such as Beijing and Shanghai is the most desirable for university graduates from outside the cities but is harder to get. In these cities, a *Hukou* is closely tied to social insurance benefits and access to affordable housing and local public schools.
educational quality involves investigating both their intended and unintended consequences (Kane, 2006, 2013; Shulha & Wilson, 2009). Thus the following research questions were formulated to guide the investigations within this study:

RQ1. What are the intended consequences of the CET-4 and what are the major procedures taken by the test developer to achieve the intended consequences?

RQ2. What are the actual uses of the CET-4 in the educational, business, and government contexts and what are the value implications underlying the actual uses?

RQ3. How does the CET-4 influence students’ learning and learning outcomes?

RQ1 concerns the test developer’s intended consequences of implementing the CET-4 program. R2 and R3 focus on the extent to which the intended consequences are being realized as well as unintended consequences.

1.2 Conceptualizations of Consequences, Impact, and Washback

The CET-4 is an English language testing program implemented for the purpose of enhancing the quality of tertiary EFL education in China. Kane (2006) argues that a test used to implement education policy should be evaluated in terms of its consequences. Evaluation of a program has to address both goal attainment and system integration issues. Goal attainment refers to the question if the program can achieve goals established within it. System integration concerns how the program functions in the context in which it is implemented (Chen & Garbe, 2011). This entails uncovering the unintended effects of implementing the program (Hansen, 2005). Shulha and Wilson (2009) suggest using goal-free evaluation (GFE) in evaluating large-scale assessment programs. GFE “places an emphasis on assessing the consequences of implementing the program and the degree
to which the program is serving a profile of demonstrated needs” (Shulha & Wilson, 2009, p. 118).

In light of the above insights from program evaluation and an argument-based validation approach, consequences are defined in a broad sense in this study. Specifically, the consequences investigated in the study include the CET-4 developer’s intended goals of implementing the program; washback of the CET-4; actual uses of the test in different contexts, which imply an impact on students; and value implications underlying test uses. In short, these consequences include both the intended outcomes of implementing the CET-4 program and the functioning of the program in the context of this study.

*Impact* and *washback* are terms often used in applied linguistics and language assessment research. Impact refers to the influences of testing on individuals, groups, institutions, and society at large (Chalhoub-Deville, 2009; Cheng, 2008a). Since this dissertation draws heavily on arguments from the educational assessment literature, where the term *consequences* is commonly used, I use impact and consequences interchangeably. Washback is considered a dimension of impact. It refers to the influence of testing on teaching and learning (Wall, 1997). Both impact and washback are well-established areas of research in language testing (e.g., Cheng, Sun, & Ma, 2015).

**1.3 Context of This Study**

*The Program Evaluation Standards* (Yarbrough, Shulha, Hopson, & Caruthers, 2011) calls for evaluations to clarify exactly the program that is being evaluated and how it is contextualized. Therefore, in this section, I provide pertinent information about the CET-4 program as well as relevant information about the Chinese educational and sociopolitical contexts in which it is implemented.
1.3.1 The College English Test

The CET was launched in the mid 1980s by Chinese Ministry of Education (CMoE) to measure students’ English language proficiency in the context of EFL teaching and learning in Chinese universities and colleges. EFL education in this context is referred to as College English. It is mandated to non-English major undergraduate students in Chinese universities and colleges. College English began to be offered in the mid 1980s across higher education institutions in China. Now it constitutes a compulsory course unit for the first four academic terms (two academic years), accounting for 10% of the total credit points required for the non-English major undergraduate to obtain a Bachelor’s degree. To provide guidelines for College English education, the CMoE issued a nationally unified curriculum and has revised it over time. The current curriculum, which is referred to as the College English Curriculum Requirements (CECR), was updated in 2007 (CMoE, 2007a). Before that, the curriculum was called College English Teaching Syllabus.

The stated purposes of the CET are to examine the English proficiency of undergraduate students in China and to ensure that the students reach the required English levels specified in the College English curriculum (i.e., the CECR) (Syllabus for College English Test, 2006). The CMoE is primarily responsible for making policies related to the test requirements and test uses, and provides guidelines for the reform of the CET. It also appoints members of the National College English Testing Committee (NCETC). The NCETC is mainly responsible for the technical aspects of the test.

The CET consists of three tests: CET Band 4 (CET-4), CET Band 6 (CET-6), and CET-Spoken English Test (CET-SET). Students, usually second-year students who have
completed the College English courses Band 1 to 4, take the CET-4. Students who have completed the College English courses Band 5 to Band 6, usually third-year students, take the CET-6. The CET-4 and the CET-6 are held twice a year in June and December at the end of each semester. The CET-SET was introduced in 1999 and is also held twice a year. To be eligible to take this test, students need to write the CET-4 and obtain a score over 425. This thesis study focuses on the CET-4 because: 1) College English courses Band 5 and Band 6 are not mandatory in most Chinese English universities and colleges; 2) not all the students are eligible to take the CET-SET; and 3) discussions on the CET impact have generally focused on the CET-4 both in academic literature and social media.

Scores of each administration of the CET-4 are equated first using anchored students and then are normalized against a pre-determined norm group. The norm group for the CET-4 is made up of approximately 30,000 non-English major undergraduates from 16 Chinese universities. In the current version of the CET-4, the total score ranges from 220 to 710, with a mean of 500 and a standard deviation of 70. In the early stage of the test, two scores were reported: the percentage score and the percentile score. The percentage score was reported on a 100-point-scale basis, with 60% as a passing score. Certificates were issued by the CMoE to students who passed the test.

The structure of the CET-4 and the time for the test administration are presented in Table 1.1.
### Table 1.1

**Structure of the CET-4 Test Paper and Administration Time**

<table>
<thead>
<tr>
<th>Test component</th>
<th>Question type</th>
<th>Number of questions</th>
<th>Points (out of 100)</th>
<th>Time (in minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part I. Writing</td>
<td>Passage writing</td>
<td>1</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Part II. Skimming and Scanning</td>
<td>Multiple choice and short answer</td>
<td>10</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Part III. Listening Comprehension</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section A Short Conversations</td>
<td>Multiple choice</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Long Conversations</td>
<td>Multiple choice</td>
<td>7</td>
<td>7</td>
<td>35</td>
</tr>
<tr>
<td>Section B Listening Passages</td>
<td>Multiple choice</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Section C Compound Dictation</td>
<td>Filling in blanks</td>
<td>11</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Part IV. Reading in Depth</td>
<td></td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Section A Banked Cloze</td>
<td>Filling in blanks with given words</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Section B Passage Reading</td>
<td>Multiple choice</td>
<td>10</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Part V. Cloze</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part VI. Translation</td>
<td>Chinese to English sentence translation</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>92</td>
<td>100</td>
<td>130</td>
</tr>
</tbody>
</table>

### 1.3.2 Context of the CET Implementation

The focus of this study is on the consequences associated with the interpretations and uses of the CET-4. These consequences need to be understood in relation to the context in which the test is being implemented. The effects of testing practices accumulate to help construct the social reality, which in turn influences the ways that scores of a particular assessment are interpreted and used (Moss, Girard, & Haniford, 2006). Two contextual factors are relevant to interpretations and uses of the CET-4: 1) the long history of large-scale testing in China and its impact on Chinese education and society; and 2) the prestigious status of the English language in China in the present day.
China has a longer history of large-scale testing than anywhere else in the world. The imperial civil service exam system (called Keju or 科舉 in Chinese) in Han dynasty (206 BC to AD 220) is the earliest known, formal large-scale test (Cheng, 2008b; O’Sullivan, 2012). It served the purpose of selecting officials for the state based on merit and education, with the added benefits of instituting meritocracy and maximizing political stability (Suen & Wu, 2006). However, its detrimental impact on test takers and the educational system emerged later on. Suen and Yu (2006) summarized the long-lasting chronic negative consequences of Keju: memorizing by rote of model performances; focusing on test-taking skills and surface features; cheating; and psychopathological effects. Han and Yang (2001) argue that the system strengthened the utilitarian values of education and schooling, stressed the key role of examinations in education, stressed book knowledge at the expense of practical ability, and emphasized the one-off result of testing to the neglect of formative assessment. O’Sullivan (2012) points out the Keju system is possibly the best example of negative washback as the test itself led to its own education industry and the education system focused solely on preparing people for passing the exams. In the end, “the education system had been completely destroyed by the test” (O’Sullivan, 2012, pp. 9-10).

Due to these harmful effects, the Keju system was abolished in 1905. However, its influences on essential aspects of Chinese society and education can still be felt (Chen, 2011; Cheng, 2009). It still greatly influences Chinese people’s ideology about education and schooling (Han & Yang, 2001). Today, meritocracy is still an ideology with powerful political attraction in China (Bell, 2008). The current Chinese educational system is still an exam-driven selection mechanism that measures and rewards merit (Liu, 2013).
The National College Entrance Examinations (NCEE) (or Gaokao, 高考 in Chinese) is called by many Keju in the modern society. It is a national university admission testing system including three exams: Chinese language, mathematics, and foreign language (mostly English). Each year, around 10 million students take the NCEE. Taking the NCEE is considered the only fair chance for a good future for most secondary graduates, especially those from unprivileged families (Liu, 2012). The examination scores determine if a student will be able to attend a university and which university he/she is able to attend. The university a student attends sets the course of his/her life as higher education is still the most important way of enhancing social status for the majority of the population in China (Han & Yang, 2001).

Associated with the high stakes of the NCEE are heated debates over its consequences. On the one hand, there is evidence that the intended outcome of selection is achieved from the perspective of the correlation between students’ NCEE scores and their GPAs at universities (e.g., Wu, 2007). On the other hand, there is evidence showing the negative washback and adverse social consequences of the NCEE. For example, Wong (2012) argues that the NCEE promotes rote learning and fetters creativity. It leads to enormous psychological strain on students and favors students from large cities and well-off families. Serious social consequences such as suicide and cheating are frequently reported (http://www.time.com/time/world/article/0,8599,1631854,00.html; http://news.xinhuanet.com/english2010/china/2011-06/06/c_13913171.htm).

The overview above highlights the tremendous impact of large-scale high-stakes examinations in China on individuals, the educational system as well as society at large. Cumulatively, such examinations contributed to the construction of the Chinese social
reality. Despite the growing concerns about their consequences, examinations continue to enjoy societal acceptance and recognition in China today as a fair means for selection and social mobility (Cheng, 2009). The interpretations and uses of the CET-4 as well as the associated consequences are situated within this context.

With regard to the prestigious status of English in Chinese society, EFL was given special importance in China as an immediate result of its open-door policy in late 1970s. Promoting EFL education was an important initiative in the Chinese government’s opening-up policy (Lam, 2005). It was expected that with the high English proficiency standards for university undergraduates, China could have more exchanges with the outside world and could catch up with the developed nations more quickly (CMoE, 1986; Hu, 2005). Recently, assistant Minister of Ministry of Foreign Affairs of China, Le Yucheng, made the following remarks to illustrate the attainment of this goal:

Here are two interesting statistics. One is that 400 million Chinese people have been lifted out of poverty over the past 30 years. The other is that 400 million Chinese have learned English over the past 30 years. At the first, it might seem that the two figures are unrelated. But I believe there are close links between the two. Without learning from the West, we could not have raised so many people out of poverty, at least not so fast.

—“Different Kind of Exceptionalism”, China Daily, 24 June 2011

In the educational context, English is a compulsory course taught at the primary, secondary, and postsecondary levels in China. The importance of English to Chinese society is emphasized in the English curricula at these levels. Various EFL tests at the school, municipal, provincial, and national levels are being used as gate-keeping devices
(Wang, 2007). Passing these tests is regarded as the key to success (Cheng, 2008b) and as a passport to better academic, professional, and social positions (Hu, 2005). Among these tests, the CET is the largest in scale and has the most intense impact in College English education, on individual students, and on society at large.

Concerns over the CET’s impact have generated debate over its positive and negative consequences in both the educational and the societal contexts in China (Cheng, 2008b). Washback of the CET has received substantial research attention. Studies continue to report mixed results (I will review these studies in Section 2.4.2). Voices suggesting abolishing the CET have even been heard (Ma, 2014). In this context, the CET developer is paying increasing attention to consequences of the test. Wu Qidi, vice minister of the CMoE, summarized from the policymaker’s perspective the positive impact of the CET as follows:

The fact that such a large-scale test has been developing steadily in the past few decades is in itself solid evidence to show that the CET has met the social needs, won social recognition, produced beneficial effects on society and contributed significantly to the continued improvement of the quality of College English teaching in China (quoted from Jin, 2009, p. 55).

The above overview indicates the social values associated with EFL in the Chinese context as well as the power of EFL tests, particularly the CET. For these reasons, evaluation of the CET-4 consequences goes beyond the measurement level to the policy level.
1.4 Rationale of the Study

The decision of using a testing program as a policy instrument for educational reform is necessarily evaluated in terms of consequences because the criteria for evaluating any policy or practice are the consequences associated with the policy or practice (Kane, 2012). Various consequences—intended and unintended, positive and negative, direct and indirect—are associated with the implementation of the CET-4. Therefore, my major methodological considerations in this study were: 1) what evidence to be collected; and 2) how to synthesize the evidence collected to make my evaluative arguments.

With regard to the first consideration, the test developer (also policymaker in the CET context), test users, and students are the three major players on the CET-4 scene. They constitute the sources for data collection in this study.

First, data concerning the CET-4 developer’s intended interpretations and uses of the test provide a deep understanding of the intended consequences of implementing the CET-4 program. These also constitute baseline evidence against which evidence from other sources is discussed to explore the extent to which the intended consequences are being achieved. Data regarding the CET-4 developer’s efforts in support of the intended interpretations and uses shed light on understanding how the intended consequences are to be achieved from the test developer’s perspective.

Second, consequences, particularly the high-stakes impact on students, are associated with actual test uses. Therefore, in order to unpack the impact of the CET-4, data concerning how the test results are actually used in context should be collected. Investigations of the actual uses also shed light on understanding washback of the test on
students’ learning, which is a mediated activity that occurs in an interaction between the individual and the social environment from the sociocultural and self-regulated learning (SRL) perspectives (Pintrich, 2000; Rose, 2012; Zimmerman & Kitsantas, 2005). Actual uses of the CET-4 constitute an important part of the learning environment. Investigation of actual uses of the CET-4 entails investigating value implications underlying test users’ interpretations of the test. This is particularly important in understanding how the CET-4 is used to serve the demonstrated needs of test users in different contexts.

Third, considering that the CET-4 is a testing program for the purpose of enhancing education quality, evidence about washback on students’ learning and learning outcomes is a legitimate, necessary, and crucial part of the test evaluation. In this study, I link the findings concerning the CET-4 washback to findings regarding the intended and actual uses of the test as well as the test developer’s efforts in support of the intended interpretations and uses. The purpose of this linking is to address the research problem: why the CET-4 washback exists in the Chinese tertiary education context.

With regard to the synthesis of evidence in this study, I find the argument-based validation approach particularly useful. This approach provides a methodological guideline for collecting, synthesizing, and discussing evidence from different sources. This synthesis allows me to make evaluative arguments about the decision of using the CET-4 testing program for educational reform in China. I will explain how this approach is employed in this study in the following Chapter (see Section 2.2).

1.5 Significance

This study is a timely and original attempt to investigate the consequences of the CET-4 within the argument-based validation framework. Linking the CET-4
consequences with the interpretations of its construct by the test developer and the local test users, this study not only provides evidence as to what CET-4 washback looks like and what brings about the washback, but also explains why the washback exists in the context of this study. It contributes to language assessment and educational assessment and has significant practical implications for the CET.

First, this study makes significant contributions to washback research. Previous washback studies in the field of language assessment have been conducted independently of validity (Bachman, 2005; Cheng, 2008a). This study fills this research gap by connecting washback evidence to evidence from other sources as mentioned above in a coherent validation framework.

Second, this study takes a broader view of test validity encompassing the uses of the CET-4 and the values inherent in the test construct interpretations and uses. It informs a social perspective on language testing, placing it beyond the micro level of testing construct into the macro social context by investigating values implicit in the CET-4 implementation.

Third, the study investigates consequences of the CET-4 by involving perspectives of both the test developer and test users. Findings of this study have significant implications for partitioning responsibilities between the test developer and test users, an issue that has received increased attention from researchers in the field of educational assessment since the 1990s (Nichols & Williams, 2009).

Fourth, findings of this study concerning the effects of the CET-4 on students’ learning and their learning outcomes as well as the uses of the CET-4 test score in different educational and societal contexts shed light on understanding validity of the
CET-4 in the Chinese context. These findings have significant practical implications for the CET-4 practice in terms of enhancing its positive consequences and reducing its negative consequences.

### 1.6 Dissertation Structure

The dissertation consists of six chapters. Its structure is illustrated in figure 1-1.

![Figure 1-1. Structure of the dissertation](image)
In this chapter, I first presented the purposes of the study and articulated the specific research questions. Then, I conceptualized three major terms in the study: consequence, impact, and washback. Following this, I introduced the research context before providing a rationale for the study. Finally, I summarized the significance of the study in terms of its contributions to language testing research and its practical implications for the CET-4.

In the following chapter, I first review contemporary approaches to validity, particularly the argument-based approach, focusing on how test consequences have been discussed in these approaches. Then, I explain how the argument-based validation approach is employed in this study. Finally, I review research into consequences, impact and washback in both educational and language assessment literature. Based on this review, I highlight the washback research gap in language assessment and the research problem this study addresses.

The third chapter focuses on methodology. First, the multi-phase multi-method research design of the study is described. Then, my role as the researcher in the study is clarified. Finally, details are provided about participant sampling, instruments for data collection, the data collection procedures, and data analysis methods.

The fourth chapter presents results about 1) intended consequences of the CET-4; 2) major procedures taken by the developer to achieve the intended consequences; 3) actual uses of the CET-4 results in the educational, business and government contexts; and 4) value implications underlying the CET-4 interpretations and uses.

The fifth chapter reports results of the structural equation modeling SEM analysis regarding the washback of the CET-4 on students’ learning and learning outcomes.
The sixth chapter discusses the results reported in Chapters 4 and 5, and draws conclusions based on the findings to address the three research questions. It also summarizes the contributions as well as limitations of the study and points out the implications of the findings for the CET-4 practice and for future research.
Chapter 2 Literature Review

To address the research problem “why the CET-4 washback exists in the Chinese educational context”, this study evaluates the decision of using the CET-4 program as a policy instrument for educational reform. Drawing on the argument-based validation approach and insights from program evaluation, it focuses on the consequences of the CET-4 program in the Chinese educational and societal contexts, particularly its washback effects on students’ learning and learning outcomes. This chapter reviews literature on validity and consequences in educational assessment and language assessment. It consists of five sections. Section 2.1 reviews literature on validity, focusing on the role of consequences in test validation, particularly how this is discussed in the argument-based approach. Following this review, in Section 2.2, I explain how the argument-based approach is used in this study to develop a framework for the intended interpretations and intended uses of the CET-4 score from the test developer’s perspective. This framework provides a methodological guideline for this study. Section 2.3 reviews research on consequences in educational assessment literature. Section 2.4 reviews impact and washback studies in language assessment literature, including literature on washback of the CET-4. Section 2.5 summarizes this chapter and highlights the research gap and the research problem this study addresses.

2.1 Test Validation and Consequences

The trajectory of major developments in validity theory over the past few decades is that the terms validity and validation have been consistently defined in terms of the interpretations and uses of test scores rather than a property of the test (American
Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1966, 1974, 1985, 1999; 2014; Cronbach, 1971; Cureton, 1951; Kane, 2006; Messick, 1989). However, there are contrasting views regarding the kinds of interpretations and uses to be considered in validation, and consequently divergent strands in the literature regarding whether and how test consequences should be included in validation inquiries.

One strand gives priority to score interpretation over score use. Researchers in this strand (e.g., Borsboom, Mellenbergh & Van Heerden, 2004; Cizek, 2012; Mehrens, 1997; Popham, 1997; Reckase, 1998) attempt to simplify validation by exclusively focusing on interpretation as the basis for validation. For example, Mehrens (1997) argues “one can investigate the validity of the inference that a score is a reasonable indicator of the amount of a construct possessed independent of any specific use of the score” (p. 17). Later, Lissitz and Samuelsen (2007) suggest that validity should be concerned with internal test construction processes with an emphasis on reliability and content representation. They assert that consequences associated with test uses emerge from complex contexts, and so considerations of consequences will overburden the conception of validity. More recently, Cizek (2012) argues that the measurement community could accept the responsibility of investigating intended consequences and major sources of unintended consequences but address these issues under some other heading such as ethics. This strand highlights the concerns of the measurement community over the complexity of test consequences and over the challenges of identifying, specifying and evaluating all the potential consequences in precise quantitative terms.
In contrast, other researchers focus on test use and consequences in conceptualizing validity. For example, in the first “Validity Chapter” in *Educational Measurement*, Cureton (1951, p. 621) defined validity in terms of the relevance of the test to its intended uses. “The essential question of test validity is how well a test does the job it is employed to do” (Cureton, 1951, p. 621). Later, Cronbach (1980, 1988, 1989) provided insights gained from program evaluation, arguing that validation of test use is evaluation and evaluation involves much more than a determination of *truth* or *accuracy*. It also involves arguments and judgments of *worth* that demand attention to consequences of test uses and interpretations. “Tests that impinge on the rights and life chances of individuals are inherently disputable” (Cronbach, 1988, p. 6). More recently, Moss (1998; Moss et al., 2006) took a further step, considering the relationship between validity and consequences from a hermeneutics perspective. She highlights the dialectic relationship between assessment and social reality, arguing that the study of consequences is an essential aspect of validity even for those who choose to limit the scope of validity to a test-based interpretation. Practices of testing change the social reality, which in turn, shapes interpretations of test construct and test uses. Therefore, test consequences need to be considered and investigated in relation to the issues of ethics and power that shape the consequences. This assertion points to the importance of revealing the social and political values implicit in test constructs and test implementation.

Messick (1989) proposed a four-faceted conception of validity to integrate test interpretation and test use within one framework (Table 2.1).
Table 2.1

*Unitary Validity Model*

<table>
<thead>
<tr>
<th>EVIDENTIAL BASIS</th>
<th>FACETS OF VALIDITY</th>
<th>CONSEQUENTIAL BASIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct validity</td>
<td>Construct validity + Relevance/utility</td>
<td></td>
</tr>
<tr>
<td>Value implications</td>
<td>Social consequences</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Adapted from Messick (1989, p. 20).

The unitary validity model starts with a traditional investigation of test score meaning and then adds test relevance, values, and further social consequences. Together, these four facets form a progressive matrix. Messick’s (1989) significant contribution to the discussion of validity is the explicit inclusion of the consequential basis in his validity model (i.e., both value implications associated with test interpretation and social consequences associated with test use). However, in contrast to Cronbach (1988), who maintained that consequences could invalidate test use, even if they could not be traced to any flaw in the test, Messick (1989, p. 88) argued “if the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced ... then the validity of the test use is not overturned.” He suggests, from a test design perspective, promoting positive consequences through minimizing sources of invalidity in terms of construct under-representation and construct-irrelevant variance.

Although Messick (1989) proposed a progressive matrix for validation, he provided virtually no methodological guidance as to how to go about developing evidence for validation (i.e., where to start, how to proceed, when to stop, how to weigh different evidence and how to synthesize different evidence) (Chapelle, Enright, &
Jamieson, 2010; Kane, 2012). To address this issue, Kane (2006), following Cronbach (1988) and drawing on practical argumentation theories (Toulmin, 2003), conceptualizes validity as an argument. The argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies a network of inferences. The inference is justified by a warrant, which is a general rule for inferring claims of a certain kind from data of a certain kind. Warrants are not generally self-evident. They can be supported or challenged by evidence. Evidence supporting a warrant is its backing; evidence challenging a warrant is its rebuttal. More recently, Kane (2013) used the term “interpretation/use argument” (IUA) to refer to the network of inferences and warrants inherent in the proposed interpretation and use. Validation is evaluation of the coherence and completeness of the IUA and of the plausibility of its inferences by collecting, analyzing, and synthesizing evidence supporting or rebutting the warrants underlying the inferences. The kinds of IUAs determine the kinds and amount of evidence for validation.

In cases where tests are used to make decisions, the warrant for the decision is the decision rule. It is the capstone of the IUA, which generally involves a chain of inferences leading to claims about score interpretations and then to decisions based on the interpretations. Thus evidence for the accuracy of the score interpretation on which decisions are based is clearly relevant to the evaluation of the decision. More accurate interpretation generally should lead to more positive consequences and inaccurate interpretation would undermine the rational basis for the decision. Moreover, the use of inaccurate information as the basis for a decision is unacceptable ethically, legally, and socially. However, evidence for the validity of a test-score interpretation does not, in
itself, justify a score-based decision because the decision rule has to be evaluated in terms of consequences. For the decision rule to be justified, the intended outcomes should be achieved at a reasonable cost and with acceptable unintended consequences. However, negative consequences do not, in themselves, invalidate the score interpretation because bad decisions can be made based on sound score interpretations, but suggest a closer scrutiny of the testing program. Therefore, like Messick (1989), Kane (2006; 2013) links consequences to construct validity, and like Cronbach (1988), he maintains that consequences can invalidate test use, even if they are not traceable to any flaw in the test (i.e., they may not necessarily jeopardize the validity of interpretation).

In language assessment, Bachman (2005; Bachman & Palmer, 2010) proposes an assessment use argument (AUA) based on the argument-based approach to validation (Kane, 1992, 2001, 2002, 2004; Mislevy, 2003). The structure of the AUA is presented in Figure 2-1.

![Assessment Use Argument Diagram](image)

**Figure 2-1.** Structure of assessment use argument. Adapted from Bachman (2005, p.25).

Like Kane (2006, 2013), Bachman refers to the AUA as the overall argument linking assessment performance to use (decision). It includes two parts: a utilization
argument, linking an interpretation to a decision, and a validity argument, linking
assessment performance to an interpretation. It is noticeable, however, that Bachman is
not consistent with Kane (2006, 2013) in using terms. What Bachman refers to as the
utilization argument in the AUA is referred to as the decision rule in Kane’s (2013) IUA.
Consequence is dealt with differently by Bachman (2005) and Kane (2006). Bachman
(2005) suggests four types of warrants for a utilization argument, and intended
consequences are one of them (the other three being relevance, utility, and sufficiency),
while for Kane, the decision rule is the warrant for any policy or practice involving test
use, and consequences are the criteria for evaluating the decision rule.

For researchers, the major strength of the argument-based approach is that it
provides a transparent working framework to develop and evaluate the IUA for particular
assessment use contexts. Two recent studies demonstrated this strength. Chapelle,
Enright, and Jamieson (2008, 2010) applied the network of inferences conceptualized and
codified by Kane (2006) in developing an IUA for the Test of English as a Foreign
Language™ (TOEFL®). The argument-based approach allowed them to frame the
intended TOEFL® score interpretations and uses based on purposes of the test, outline the
essential research, and structure research results into a validity argument. In a more
recent study, Cheng and Sun (2015) evaluated the strength of the IUA for the Ontario
Secondary School Literacy Test (OSSLT) in Canada from the perspective of its impact
on second language (L2) learners. Drawing on Kane’s (2006) framework, they were able
to specify the IUA by laying out the network of inferences based on the purposes of the
OSSLT, and to synthesize empirical evidence derived from a large-scale, mixed-method
study on impact of the OSSLT to challenge the IUA.
2.2 Application of the Argument-based Validation Approach in This Study

Drawing on Kane’s (2006, 2013) argument-based validation approach, an IUA framework for the CET-4 was built in this study based on the stated purposes of the test. This framework is presented in Figure 2-2.

**Figure 2-2. An IUA Framework for the CET-4**
Figure 2-2 lays out the whole network of inferences involved in the IUA for the CET-4. On the right side of the figure, these inferences (target domain description, scoring, generalization, explanation, extrapolation, and utilization) are represented by the vertical arrows. Each inference extends the interpretations of the test score represented by the ovals (observation, observed score, expected score, construct, target score), leading to the CET-4 use decision. Presented in the boxes on the left are the warrants for each of the inferences as well as the decision rule (i.e., the warrant for the decision).

Figure 2-2 is read starting at the bottom with the target domain. The target domain of the CET-4 is defined by the College English curriculum (i.e., the CECR). Observation refers to performance of the test-takers on the CET-4 for which the interpretation is made. The target domain description inference refers to efforts to describe the domain carefully and to develop items that reflect the domain. Observed scores are test-takers’ scores on specific tasks of the CET-4 or the whole test. To obtain the observed score, each test-taker’s performance (observation) needs to be assigned a score using a scoring procedure. This process is referred to as the scoring inference. Generalization extends the interpretation of a test-taker’s observed score on one CET-4 test to the expected scores across similar assessments. The construct refers to the English language proficiency that accounts for the consistencies in individual performances. The explanation inference is the effort to explain the performance consistency in terms of construct. The extrapolation inference links the construct of English language proficiency to the target score (i.e., the quality of students’ performances in the EFL use contexts specified in the CECR). Finally, the utilization inference (i.e., the effort to make the CET-4 score relevant and
useful to test users in the tertiary education context in China) links the target score interpretation to test use decision.

At the top of this network is the decision. It is included in the IUA framework because the implementation of the CET-4 program is a policy decision. Chalhoub-Deville (2009) noted that in the context in which testing programs are used as policy instrument for educational reform, government agencies assume the roles of both test developer and test user. They “are explicitly dictating testing requirements and are increasingly shaping how scores should be interpreted and used” (Chalhoub-Deville, 2009, p. 124).

For the decision to be justified, it has to be based on accurate interpretation of the test score. Thus evidence about all the inferences in the IUA in Figure 2-2 is clearly relevant to the evaluation of the decision. The warrant for the decision is the decision rule, which is evaluated in terms of intended and unintended consequences of implementing the CET-4 program.

In developing the framework in Figure 2-2, I drew on the IUA for TOEFL® (Chapelle et al., 2008) and the IUA for the OSSLT (Cheng & Sun, 2015). However, it should be noted that although the principle and the coding for building up the IUAs for different tests may be the same, the details involved in the argument vary from test to test. Different tests target different domains, use different procedures, measure different constructs and their scores are intended for different use decisions. Therefore, the pertinent inferences and assumptions involved in the IUA for a particular test are specific to the test context and are affected by design decisions made in the context (Xi, 2008). Evidence needed for the IUA evaluation in the TOEFL® or OSSLT context is different than evidence needed for the CET-4 IUA evaluation. This IUA framework presented in
Figure 2-2 guides the research design of this study. Findings of this study are also discussed within this framework (see Chapter 6) to address the research questions.

2.3 Research into Consequences of Large-Scale Assessment

“Testing is never a neutral process and always has consequences” (Stobart, 2003, p. 140). The measurement community has a long tradition of investigating intended and unintended direct consequences (i.e., consequences associated with uses that rely directly on the information provided by scores to make decisions, particularly high-stakes decisions). Recently, the extensive uses of large-scale assessment programs for purposes such as accountability and informing as well as influencing classroom practices have generated increasing interest in the beneficial and detrimental consequences of such programs in the educational system (Rogers, 2014). The systemic effects of assessment are generally indirect consequences, which capitalize on the motivational effects of testing or use testing and test reporting to shape public opinion (Haertel, 2013).

2.3.1 Intended and Unintended Direct Consequences

Intended consequences (i.e., desired outcomes and definable benefits that are planned and anticipated by the test developer and/or test users) of direct score uses have generally been investigated under the heading of validity. As Kane (2013, p. 58) noted, “for the first half of the 20th century, validity was goal-driven, aimed at predicting (and to the extent possible, maximizing) some desired outcome.” The focus was particularly on the direct consequences of using information provided by test scores for selection purposes. For example, in the context in which a cutscore has a major role in making high-stakes decisions in employment, the cutscore is determined by analyzing the
consequences (e.g., the productivity of workers hired) of the adoption of different cutscores (Cronbach & Gleser, 1965).

Compared with intended consequences, unintended consequences are more difficult to evaluate since they “are varied, unplanned, and often harder to anticipate. Before they can be investigated and weighed, they must first be noticed or discovered” (Haertel, 2013, p. 85). Unintended consequences may arise from both the intended/legitimate and unintended/illegitimate uses of the test score. Unintended consequences associated with legitimate uses are the side-effects of the test (Shepard, 1997). Side-effects, particularly adverse consequences on test takers, have long been discussed from two closely connected perspectives: fairness and validity (e.g., Camilli, 2006; Cascio, Jacobs, & Silva, 2010; Cole & Moss, 1989; Ebel, 1966; Kane, 2010; Messick, 1989). A test is considered fair and valid to the extent that the adverse impact is not traceable to construct-irrelevant factors and/or under-representation of the test construct. Therefore, fairness considerations are often built into test developers’ efforts for reliability and validity in the large-scale assessment context. For example, if there is a constructed-response component in the assessment, rubrics are carefully developed, markers are effectively trained, and the validity and reliability of the scoring is continuously monitored so that all students’ responses are scored reliably and fairly (Alberta Education, 2012; EQAO, 2012b). Kane (2010, p.178) points out that “the whole point of standardized testing is to treat everyone in the same way.” He calls fairness in the sense of equality procedural fairness. It is more related to the quality of the test per se than to the test use.
Unintended/illegitimate uses of the score are misuses or abuses of the test. Evaluating unintended consequences goes beyond investigating procedural fairness. A test that is procedurally fair may be misused and/or abused and have harmful consequences. Therefore, Kane (2010) asserts that procedural fairness is an essential and necessary requirement for both fairness and validity in testing, but this is not enough. It is also necessary that the score interpretation and any test-based decision rule be reasonable and appropriate. He calls this substantive fairness. Substantive fairness is concerned with how the testing program functions and includes a much wider range of issues than procedural fairness. Investigating substantive fairness involves exploring values underlying test interpretations and test uses and making value judgments. Therefore, substantive fairness is more value-laden and more difficult to evaluate compared with procedural fairness.

Related to the distinction between side-effects and consequences associated with misuse/abuse of the test is the question: “Who should take the responsibility for evaluating testing consequences?” Messick (1998) and Shepard (1993, 1997) assert that test developers should be responsible for side-effects, especially if unanticipated adverse effects are traceable to sources of test invalidity, and the responsibility for misuse/abuse of the test lies with the (mis)user. Kane (2006) argues for shared responsibilities of the test developer and test users with regard to social consequences of assessment programs. While test users have the primary responsibility, in making decisions, they may rely on their interpretations of the test score and/or evidence and analyses provided by the test developer. Thus it would be reasonable to expect that the test developer would provide accurate and interpretable information to test users.
More recently, Nichols and Williams (2009) proposed a useful framework to further clarify the responsibilities of test developer and test user (Figure 2-3). This framework is organized around three dimensions: 1) breadth of construct (i.e., the extent to which the construct is defined). The broader the advertised construct representation, the more extensive is test developer’s responsibility for documenting the consequences of score use; 2) test use (i.e., the difference between the actual use of the test score and the intended use of the score by test developer). The greater the difference, the more test user must be held responsible for documenting the evidence of test score use consequences; 3) time (i.e., the amount of time since test publication). Test developers’ responsibilities may expand or shrink over time depending on how the test score is actually used. This framework also depicts an area called zone of negotiated responsibility between the test developer and the test user, in recognition that the ownership of responsibility is not always clearly delineated.

![Figure 2-3. Delineating the responsibilities of test developers and test users. Adapted from Nichols & Williams (2009, p. 7).](image-url)
In principle, Nichols and Williams’ (2009) framework is not different from previous discussions on the roles and responsibilities of the test developer and test users (e.g., Messick, 1998; Shepard, 1993; 1997). The usefulness of this framework lies in that it captures the fluid nature of validation, consequences, and responsibilities, and includes the zone of negotiated responsibility. This is particularly helpful in cases in which the distinction between the test developer and test users is blurred. This is also where researchers should play their important roles. However, Nichols and Williams (2009) did not discuss researchers’ roles in this framework.

Despite the above clarification, there are some practical issues to be addressed to partition responsibility for the evaluation of consequences between the test developer and test users. For example, if test developers are accountable for the intended effects and side-effects of the legitimate uses of their tests, there might be a confirmationist bias that dogs much of their validation practice (Cronbach, 1988; Haertel, 1999). Test developers, whether commercial test publishers or government testing authorities, may engage more energetically in gathering evidence to support reported validity of test use than in studies that might discredit their products. Test users, on the other hand, may take it for granted that the desired effect of their test use policy will be forthcoming without thinking about the need for evidence that a policy has the intended effect. They may be even less likely to push for evidence regarding unintended negative effects that their test use policy may have (Linn, 1998). Moreover, even if test users are willing to take the responsibility to conduct validity investigations, often, they are not qualified, or lack the necessary resources (Shepard, 1993).
To address these issues, Cronbach (1989) proposed the notion of independent researcher or test evaluator from the program evaluation perspective. “The evaluator holds no brief for or against the test”, but rather undertakes independent research “to serve all the persons having stakes in affairs the test may influence” (Cronbach, 1989, p. 164). Following this notion, later researchers (e.g., Chudowsky & Behuniak, 1997; Lane, Parke, & Stone, 1998; Pomplun, 1997; Ryan, 2002; Stone & Lane, 2003) identified, described, and examined unintended consequences of assessments in different contexts from the program evaluation perspective. They obtained evidence from various stakeholders—principals, teachers, students, and school administrators—using a number of methods such as surveys, interviews, focus groups, artifacts analysis, and classroom observations. Lane et al. (1998) noted that triangulation of evidence obtained through multiple measures from various levels as well as from various stakeholders has the potential to strengthen the soundness of the results about the consequential aspect of validity. More recently, Shulha and Wilson (2009) suggest evaluating large-scale assessments using an effects model of evaluation (Hansen, 2005), or goal free evaluation (Scriven, 2005), which “places an emphasis on assessing the consequences of implementing the program and the degree to which the program is serving a profile of demonstrated needs” (Shulha & Wilson, 2009, p. 118).

Chalhoub-Deville (2009) emphasizes the collaboration between researchers and policymakers in investigating potential consequences of using testing programs as policy instruments. She suggests that test developers and researchers should be active partners in the formulation of policies to better serve education. Chalhoub-Deville introduced the concept of social impact analysis (SIA) to argue for an expansion of the traditional
conceptualization of consequence research. Chalhoub-Deville discussed how some of the adverse consequences associated with No Child Left Behind policies could have been avoided had SIA been undertaken because SIA emphasizes anticipatory impact and a proactive approach to studying potential consequences before a policy is put in place.

2.3.2 Intended and Unintended Indirect Consequences in the Educational System

Since the 1990s, the educational system worldwide has become more “politically charged and value-laden” (Leach, McCormick, Moon, & Murphy, 1999, p. i), which demands “efficiency and quality” out of its expenditure (Mok, 2005). Policymakers realize that “real change will not take place in schooling until significant change happens to assessment” (Torrance, 1996, p. i). In this context, large-scale assessments are being extensively used worldwide as policy instruments or agents for educational reforms (Klinger, DeLuca, & Miller, 2008; Levin, 1998; McEwen, 1995; Nagy, 2000; Roach & Frank, 2007; Rogers, 2014; Stevenson, 1996; Wagner, 2010) because assessments can be relatively inexpensive, relatively quick to implement, externally mandated, and the results are visible (Linn, 2000). Many educational systems established new curriculum policies defining standards students are expected to achieve at specific points in educational programs and had large-scale testing programs designed to measure students’ achievement of these standards. A guiding rationale for these policies is to make the expectations explicit so that these can be taught, studied, and measured, and students’ achievements can be assessed, reported, certified, monitored, and accounted for with increased clarity, consistency, and precision (McKay, 2007).

Results of a large-scale assessment can be effectively used by educators to inform their classroom practices so that students’ learning is improved (Rogers, 2014). The
alignment between assessment, curriculum, and instruction would give a clear indication of the level of knowledge expected of students (Pellegrino, 2014; Porter & Smithson, 2000); and students will have the opportunity to learn and to truly demonstrate what they have achieved (Martone & Sireci, 2009). Assessment with feedback produces the strongest positive effect on achievement (Phelps, 2012). Such assessment results provide a means for teachers to know the effect of their instruction on learning (Anderson, 2002; Mirazchiyski, 2013; Phelps, 2012). Teachers can use the data from a large-scale assessment as an instructional diagnosis when examining strengths and weaknesses at the student and class level, which can aid in adapting programming and instruction (Goldberg & Roswell, 2000; Klinger et al., 2008). For this purpose, Rogers (2014) suggests that subtest scores from large-scale curriculum-based assessments should be reported based on the assumption that the curriculum is multidimensional. Harlen (2006) asserts that evidence collected for summative purposes can be used for formative purposes if evidence is collected frequently and provides details to inform teaching and learning, and pressure exerted by the external testing requirements is minimized.

The extent to which these consequences are being achieved is investigated by studying what is happening in schools in terms of the motivation and practices of school administrators, teachers, and students, as well as improved students’ performance (Lane & Stone, 2002; Linn, 2009). Han and Yang (2001) and Wang (2008) found that large-scale, high-stakes assessments are often used to motivate students to learn in the Chinese context. Phelps (2012) noted that adding stakes would strongly and positively affect achievement. Deci and Ryan (2002) argue that large-scale assessments are external regulators that play a positive role in motivating students, particularly students with low
intrinsic motivation. When large-scale assessments are in place, the efforts of students and teachers are more focused on the subject content, and students’ attitudes, engagement, effort, and achievement are improved (Anderson et al., 1990; Burger & Kroeger, 2003; Yeh, 2005).

However, the extensive uses of large-scale assessment in the educational system have also generated increased criticisms of its unintended detrimental effects. The criticisms have focused on “opportunity to learn” issues such as failure to test students on what they were taught and a narrowing of the curriculum because of mandated testing (Koretz & Barron, 1998; Resnick, Rothman, Slattery, & Vranek, 2004; Roach, Niebling, & Kurz, 2008). Hillock’s (2012) analysis of the evidence from the writing assessments in five states of the United States strongly suggested that these assessments promoted neither good writing nor good writing instruction. Instead, they promoted a narrow range of writing activities by teaching students to reach minimal standards. Madaus (1988) observed that in the high-stakes context, teachers would use narrow and programmatic approaches to teaching and drilling the learners. Linn, Baker, and Dunbar (1991) assert that more direct assessment will not necessarily induce classroom activities that are more beneficial to learning and that the highest stakes induce the most damaging test impact. Similarly, Crooks (1988) and Gipps (1994) argue that high-stakes standardized testing encourages shallow learning over deep learning, typically marked by the rote learning of test content without understanding. Therefore, Shohamy (1992) suggests that the stakes of testing should be minimized if schooling is to be reformed. More recently, Qi (2007) investigated the effects of the writing task in the English language test of the NMET, a high-stakes test in the secondary school context in China. She found that the high-stakes
nature of the test undermined the implementation of the intended consequences on teaching as a result of teachers’ and learners’ desires to gain high test scores and their interpretations of how best this might be achieved. She concluded that high-stakes testing was not in itself an effective means of implementing change at the classroom level.

Unintended consequences are not necessarily negative. For example, Cizek (2001) identified unintended positive consequences of large-scale testing such as professional development for educators; focusing more attention on protected kids; and promotion of teachers’ knowledge about testing. Cizek argues that it is essential to consider such consequences and incorporate them in evaluation of policy decisions.

Compared with direct uses and consequences, indirect uses and consequences, particularly unintended ones, are more difficult to evaluate (Haertel, 2013) because they are often a function of the test and other factors and forces operating in the educational system (Messick, 1996). To make it more complex, results from a single administration of an assessment are often used for more than one purpose and there are interactions among the multiple uses (Koch, 2013). To address this complexity, Koch and DeLuca (2012) propose an interpretive approach to validation drawing on the work of Moss and others (Mislevy, Moss, & Gee 2009; Moss 1996; Moss et al., 2006). This approach uses hermeneutic methodology to construct and analyze validity evidence.

2.4 Impact and Washback Research in Language Assessment

Although educational assessment has a long history of research into testing consequences, this topic was not seriously discussed in the field of language assessment until the early 1990s. Language testing researchers usually use the term *impact* to refer to the influences of testing on individuals, groups, institutions, and society at large. One
dimension of impact that has received particular attention is *washback*. It refers to the influences of testing on different aspects of teaching and learning. The concept of washback is rooted in the notion that tests can and should drive teaching and hence learning (Cheng & Curtis, 2004). Impact and washback are well-established areas of research in language testing, and studies in these areas proliferated over the past two and half decades (e.g., Cheng, Sun, & Ma, 2015).

### 2.4.1 Impact Research in Language Assessment

Messick’s (1989) validity model is most influential in language assessment validation research (McNamara, 2006). The notions of value implications and social consequences in the model are particularly welcomed in language testing, given the social nature of this practice. As Bachman (1990) pointed out, language “tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or society at large” (p. 279).

Language testing researchers have approached test impact primarily from two different perspectives: validity and ethics.

From the validity perspective, impact is considered as a quality that needs to be designed into the test. Bachman and Palmer (1996) include impact under an overarching notion of *test usefulness*, which comprises qualities such as reliability, construct validity, authenticity, interactivity, and impact. They argue that the overall usefulness of a test is a function of these qualities. The test developer needs to design them into the test and prioritize them based on the practicality of the testing situation (i.e., the kinds and amount of resources that are required). This formulation clearly brings considerations of validity and impact under a unitary concept of test usefulness. However, it does not indicate how,
if at all, these qualities are related to each other, or how construct validity and test use are directly related.

More recently, Bachman (2005; Bachman & Palmer, 2010) proposes an assessment use argument (AUA) framework (see Figure 2-1 in Section 2.1). Within this framework, intended consequences are included as one type of warrant for the utilization argument, which is based on argument for construct validity. Drawing on the argument-based approach to validation, Bachman was able to link consequences of test use to construct validity in the AUA framework. In a recent validation study, Doe (2013) drew on this framework to investigate the appropriateness of the Canadian Academic English Language Diagnostic (CAELD) assessment for diagnostic purposes in an English for Academic Purposes (EAP) program. The investigation focused on the decisions related to scoring, teaching, and learning and the extent to which these processes contributed to EAP education.

Unintended social consequences are not included in the AUA. Bachman (2005, p. 28) suggests that attention to the unintended adverse social consequences should be “incorporated into a code of ethics or a code of fair testing practice, rather than being articulated in an assessment use argument for a particular test.” This suggestion is consistent with Cizek (2012) (see Section 2.1), but contrasts with Kane’s (2006, 2012, 2013) argument that social consequences—intended and unintended, positive and negative—cannot be ignored in validation, because the criteria for evaluating any policy or practice are the overall consequences associated with them.

In contrast to the validity perspective, the ethics perspective to impact research focuses on the sociopolitical dimension of language testing. Researchers taking this
perspective (e.g., Davies, 2004; Elder, 1997; Hamp-Lyons, 1997; Kunnan, 2000a, 2000b, 2004, McNamara & Roever, 2006; McNamara & Ryan, 2011; Shohamy 2001, 2006; Spolsky, 1981, 1995) argue that language testing is not neutral and language testers need to develop a critical view of tests to critique the values underlying test construct and test use. Although these researchers use different terms and sometimes use the same term in differing and/or inconsistent ways, their interest has generally focused on linking impact and validity to fairness and/or justice.

Fairness is a particularly acute issue in language testing because of its high-stakes impact on test takers—admission, employment, residency, and citizenship decisions are often made based on language test scores (McNamara & Ryan, 2011). Decisions based on inaccurate test score interpretations are unacceptable ethically and socially. Some researchers (e.g., Kunnan, 2004) link validity to fairness and invalidity to bias. For example:

A test ought to be fair to all test takers; that is, there is a presumption of treating each person with equal respect….A test ought to have comparable construct validity in terms of its test-score interpretation for all test-takers….A test ought not to be biased against any test-taker groups, in particular by assessing construct-irrelevant matters. (Kunnan, 2004, p. 33)

This discussion is relevant to procedural fairness (Kane, 2010). The past few decades have witnessed remarkable progress of research in this area. For example, sophisticated psychometric techniques such as DIF analysis and Rasch models are commonly employed in language testing (for an overview, see McNamara & Roever, 2006, chap. 4). The purpose is to identify test items that unfairly disadvantage particular
test taker subgroups and the potential bias associated with raters that threatens reliability of the test score.

Other researchers (e.g., Fulcher, 2009; McNamara & Roever, 2006; McNamara & Ryan, 2011; Shohamy, 2001, 2007) take a broader view of fairness. They focus on the sociopolitical functions language assessment plays in the service of social policy and argue for the social responsibility of language testing researchers and language testers. For example, Shohamy (2001) calls for examining the hidden agendas of language tests from a critical view of language testing, arguing that language testing is characterized as the exercise of power by one party over another. She alerts language testers to the political uses and potential abuses of language tests and articulates the need to address issues related to the rationale for giving test and the impact of tests on test takers, education, and society.

More recently, Shohamy (2007) demonstrates how language tests play a major role in the introduction and implementation of language policy and language education policy. Language policy refers to decisions made by policymakers about languages, including foreign languages, and their uses in society. Language education policy includes such decisions about which language(s) should be taught, when (at what age), for how long (number of years and hours of study), by whom (who is qualified to teach), for whom (who is entitled and/or obligated to learn), and how (which teaching methods, curriculum, materials, tests to be used). Language policies are often stated explicitly through official documents, such as language standards, curricula, and tests. The introduction of a language test delivers messages and ideologies about the prestige and
priority of the language. Language tests serve as mediators between ideology and practice and create de facto language policies. This relationship is illustrated in Figure 2-4.

![Figure 2-4. Mechanisms affecting language policy. Adapted from Shohamy (2007, p. 121).](image)

As shown in Figure 2-4, a set of mechanisms can lead to de facto language policies. In this framework, language testing is considered a mechanism that serves as a mediator to create, enforce and perpetuate de facto language policies that interact between ideology and practice. This consideration implies that the test construct is dictated as a function of policy and that there are social and political values implicit in both test construct and intended test use.

It is in light of the critical view of language testing that McNamara and Ryan (2011) propose the distinction between *fairness* and *justice*. Fairness is defined in terms of a specific technical issue (i.e., the extent to which the test quality, especially its psychometric quality, ensures procedural equality for individual and subgroups of test-takers). Justice refers to the defensibility of the values embodied in the test (i.e., values implicit in test constructs and the social uses to which language tests may be put). The concern for fairness corresponds to the evidential basis of test score interpretation and test use, the upper row of Messick’s (1989) validity matrix (see Table 2.1). Justice is concerned with social values and social consequences, and thus, related to the
consequential basis of validity in the matrix. This distinction between fairness and justice also resembles the distinction between procedural fairness and substantive fairness made by Kane (2010). However, McNamara and Ryan’s definition of justice encompasses broader concerns about social values, while substantive fairness is more about the appropriateness of test use, particularly across different test taker groups.

McNamara and Ryan (2011) pointed out that in general, language testing has emphasized fairness from the perspective of the quality of testing procedures. The concerns for fairness and the increasingly sophisticated techniques used to guarantee it have the paradoxical potential to cloud issues of justice. Therefore, echoing Spolsky (1995), McNamara and Ryan (2011) call for a more socially aware approach to investigating the impact of language testing, taking full account of the institutional and/or political context where the test is functioning. This is also consistent with Moss’ (1998, see also Moss et al, 2006) call for situating validity argument in the sociocultural context to address questions of social consequences.

The critical ethical and sociopolitical perspective of impact research presents methodological challenges as it gives prominence to value implications underlying test interpretation and test use. The challenges lie in that “the social values expressed in test use are often implicit, so that they are not explicitly acknowledged at all and are not even debated” (McNamara & Ryan, 2011, p. 162).

2.4.2 Washback Research in Language Assessment

The early traditional view of washback was based on the general assumption about the relationship between tests and their impact (i.e., a high-stakes test would produce a washback effect, and the direction of washback is determined by the qualities
of the test). For example, Heaton (1990) asserted that a good examination will have a useful effect on teaching whereas a bad test will have a damaging effect. Early research on washback focused on the positive influences of revised well-constructed tests (Pearson, 1988; Popham, 1987) and the negative influences of multiple-choice, large-scale standardized tests on the quality of teaching and learning (Shepard, 1990).

Alderson and Wall’s work (Alderson & Wall, 1993; Wall & Alderson, 1993) was the first attempt in applied linguistics to critically look at the relationship between tests and their washback. In an article entitled ‘Does washback exist?’, Alderson and Wall (1993) questioned if washback could be a property of test validity. Based on a thorough analysis of the notion of washback, the authors suggested through a series of ‘washback hypotheses’, the sorts of areas that could be influenced by tests (e.g., what teachers teach, how they teach, what learners learn, how they learn, the rate and sequence of teaching and learning, the degree and depth of teaching and learning, and attitudes toward the content or methods of teaching and learning). Alderson and Wall (1993) argued that researchers should be precise about what they are looking for when they set out to investigate if washback has been created. Alderson and Wall’s “washback hypotheses” marked a significant development in shaping the constructs of washback studies for the field of language testing (Cheng, 2008a).

Hughes (1993) proposed a trichotomy into participants, process, and products in washback studies, noting that the nature of a test may first influence participants’ perceptions and attitudes towards their teaching and learning tasks, which, in turn, will affect teaching and learning processes, and that teaching and learning processes will affect learning outcomes. Combining this trichotomy and Alderson and Wall’s (1993)
washback hypotheses, Bailey (1996) suggests a basic washback model (Figure 2-5) depicting how a test may influence different participants engaging in different processes, and result in products specific to each category. These products may influence each other directly or indirectly. Bailey (1996) distinguishes between washback to the learners (i.e., the direct effects of the test on the test-takers) and washback to the program (i.e., the influences of the test on teachers, administrators, and curriculum developers, etc.). This washback model provided a framework for washback research in the decade to come.

![Figure 2-5. A basic model of washback. Adapted from Bailey (1996, p. 264).](image)

Over the past two decades, there has been a flurry of empirical studies investigating influences of testing on different categories of participants, processes, and products included in Bailey’s (1996) model. These studies responded to the question “what does washback look like?” —a step further from the question “does washback exist?” posed by Alderson and Wall (1993). They showed that washback is a highly complex phenomenon. Simply changing the contents or methods of an examination will not necessarily bring about the desired changes in teaching and learning. A multitude of social, cultural, and educational factors other than the test per se may have influences on
the direction (positive or negative) and intensity (strong or weak) of washback (Watanabe, 2004). Therefore, it is necessary to consider the interactions of the characteristics of the educational context, the test, the curriculum, the teacher, and the learner in explaining why washback exists in certain contexts.

In recent years, washback on students’ learning has received increased attention among language assessment researchers. In general, these studies focused on influences of learners’ perceptions and expectations of an assessment on their approaches to learning and their learning outcomes. Researchers have attempted to explain washback from perspectives such as sociocultural theory and self-regulated learning (SRL). These perspectives construe learning as the interaction between learners and their environments. “Learning is a mediated activity in which cultural artifacts have a crucial role” (Gardener, 2006, p. 57). A learner’s perception of the learning environment determines how s/he learns (Entwistle, 1991). Test preparation is a SRL process, where students’ motivation interacts with their strategic and cognitive behaviours. This process is guided by goals and contextual factors (Pintrich, 2000) and should be better understood from the sociocultural perspective (Rose, 2012; Zimmerman & Kitsantas, 2005).

Research into washback on learning highlights that students’ perceptions of test demands and test uses as well as their expectations of test-taking are crucial factors influencing learning and learning outcomes. For example, Sato and Ikeda (2015) found that students’ perceptions of multiple-choice questions (MCQs) were associated with unintended washback. They argued that directly assessing writing is necessary in order to enhance intended washback. Green (2007) proposed a model that incorporates washback direction and washback intensity (Figure 2-6). The direction of washback is determined
through participants’ characteristics and values and by the characteristics of test design. The more the test design overlaps with the focal point, the more likely it is that positive washback will be generated. Washback intensity is the result of the interaction of participants’ perceptions of test importance and test difficulty.

**Washback direction**

![Diagram of washback direction](image)

**Washback variability**

- Participant characteristics and values
  - Knowledge understanding of test demands
  - Resources to meet test demands
  - Acceptance of test demands

**Washback intensity**

- Perception of test performance
  - Important
  - Easy
  - No washback
  - Unimportant
  - Unachievable
  - Intense washback

*Figure 2-6. Model of washback incorporating direction and intensity. Adapted from Green (2007, p. 24).*

Using this model, Green (2007) investigated if a test preparation class gave learners an advantage in improving their writing test scores on the International English Language Testing System (IELTS). He compared an IELTS preparation course with university language courses. Findings of this study indicated no clear advantage for
focused test preparation in terms of increasing test scores. The study concluded that individual learners’ goals and their understanding of test demands influenced their learning outcomes to a greater extent than did their choice of course and its content. This conclusion is consistent with Perrone’s (2010, 2011) recent study, which found that students’ perceptions of the assessment task had a differential impact on their learning.

Task perception constitutes task knowledge, which plays a crucial role in effectively planning, monitoring, and evaluating the learning process (Wenden, 1998).

Shih (2007) investigated the washback of an EFL test on English learning in Taiwan. Based on findings of this study, he proposed a washback model of students’ learning (Figure 2-7), which includes influences of various factors related to the context, the student, and the test.

Figure 2-7 A Washback Model of Student Learning. Adapted from (Shih, 2007, p. 151).
Impact and washback of the CET-4 began to be reported in the late 1990s. A considerable number of empirical washback studies have reported both positive and negative effects of the College English test on teaching and learning. Its primary positive washback effects include enhancing the motivation to learn English, promoting the status of College English education in the Chinese tertiary context (Gu, 2005; Li, 2002), and improving English proficiency standards of the students in Chinese universities and colleges (Wang, 2007). The major negative washback is that it encourages test-oriented teaching and learning (Wang, 2007; Yang, 2003). In the college English classroom, more time is spent on coaching test materials, memorizing vocabulary, doing and explaining mock tests using the grammar-translation method, and developing test-wise skills rather than developing communicative competence (Gu, 2005; Jia & Yang, 2005; Ren, 2011; Wang, 2007).

The impact of the CET-4 beyond the College English classroom was reported as well. Zhang (2005) argued that the CET-4 negatively impacted students’ learning of other courses as College English was the course that took most of their time. Shu’s (2004) study showed that the CET-4 was a nightmare to some low-achievers as it had detrimental effects on their psychological well-being. In the public media, concerns over the negative consequences associated with the uses and misuses as well as abuses of the CET-4 have received increasing coverage (e.g., Guan & Ma, 2006; Li & Zhang, 2010). There have even been calls for abolishing the CET (Cai, 2005; Ma, 2014; Zhang, 2005).

Some researchers (e.g., Gu & Liu, 2005; Han, Dai & Yang, 2004; Tang, 2005; Zhang, 2003) link the negative washback of the CET-4 to its test design, pointing out that the test does not assess English communicative competence as the College English
curriculum requires due to the large proportion of MCQs. Students spend large amount of
time on learning English to pass the CET-4 (Feng, 1995; Wang, 1998, 2002) yet they
were still unable to use English effectively in communication (Jing, 1999; Liu, 2004;
Zhang, 2003). Wang and Sun (2012) identified three causes of the negative washback of
the CET-4: the ambiguous purpose of the test; the contradiction of teaching objective and
testing objective; and poor authenticity of the test. Li, Zhong, and Suen’s (2012) study
found that students adopted narrow test preparation strategies even for the direct writing
question.

Other researchers have attempted to explain the washback of the CET-4 from the
SRL and sociocultural perspectives. For example, Xie (2010) draws on the expectancy-
value theory to explain the mechanism of the washback of the test on learning. She used
SEM to model the relationships among students’ perceptions of test design and uses, their
test preparation, and test performance. Findings of this study provide empirical evidence
that both test design and test uses can affect test takers, albeit at different magnitudes and
following different paths. Xie (2010) argues that test preparation—which is dominated by
cue-seeking, narrow-focusing, rehearsal of test-taking skills, mass test paper practice,
cramming target skills, and intensive memorizing—indicates primarily negative
washback on learning in the context of the CET-4.

More recently, Xiao (2014) investigated from the SRL perspective the intensity
and direction of the CET-4 in terms of test-taking strategies students consciously
selected. This study found that students’ test-taking strategy uses were test-oriented rather
than focusing on language learning and use. The direction of the CET-4 washback on
test-taking strategy use was positive; however, the intensity was moderate or weak in terms of promoting cognitive strategy use and test management strategy use.

Chen (2011) conducted an ethnographic study with a case-study design to investigate the implementation of the CECR, focusing on recent College English assessment policy change within two Chinese universities. Participants in this study included institutional policymakers, teachers and students. The study found that excluding the CET-4 in the requirements for academic degrees within the institutional context did not remove the high-stakes that were attached to the test in the social context. Based on findings of this study, Chen (2011) asserts from a sociocultural perspective that it is not the policy initiative and the implementation that decide the assessment reality; rather, it is the local actors’ view of assessment and its nature, their understanding about the relationship of assessment and learning that are decisive factors.

2.5 Chapter Summary

Validity is at the heart of assessment evaluation and assessment research (Fulcher & Davidson, 2007). Test use decisions are evaluated in terms of consequences—intended and unintended, positive and negative, direct and indirect. The kinds of consequences considered under validity have tended to expand over time, as the understanding of how testing programs function in the world has grown. This is particularly true in the language assessment field, given the multiple educational, cultural, social, and political functions language tests have played, are playing, and continue to play worldwide. Today, they are not merely tools to measure language knowledge, but also policy instruments connected to and embedded in political, social and educational contexts.
The main concerns in validating and evaluating test use decisions are what kinds of consequences should be focused on, how the consequences are investigated, and how to synthesize and understand findings concerning consequences.

Washback of the CET-4 on students’ learning and learning outcomes is the focus of this study. Washback and impact studies are in nature program evaluation, which requires researchers to make value judgment about the local educational context as well as the larger social, political, and economic factors in relation to the testing program. Previous washback studies have unpacked the complexity of this phenomenon, but they were generally conducted independently of validity in terms of construct interpretation and use, and did not link washback findings to the macro-level context where the research was conducted (Bachman, 2005; Cheng, 2008a). Given the complexity of washback, without substantial evidence for test validation in terms of score interpretation and use in certain contexts, it would be difficult to know what washback looks like, what brings about washback, and why washback exists in a particular context (Alderson, 2004). Without answers to these questions, it will be hard to disentangle the responsibilities of the test developer and test users for the consequences of test use, particularly the unintended negative consequences for students. This is particularly relevant to the CET-4 context, given the impact of the test in the Chinese society, where large-scale high-stakes examinations have played and continue to play multiple functions.

Therefore, this study includes evidence from multiple sources and perspectives of different stakeholders (the test developer, students, and test users) in order to have a thorough understanding of the CET-4 impact on learners and their learning. It links
washback to validity to avoid the “blind elephant syndrome” (Cheng, 2008a, p. 360) in washback studies.
Chapter 3 Methodology

Drawing on Kane’s (2006) argument-based validation framework, this study evaluates the decision of using the CET-4 program as a policy instrument for educational reform, focusing on its consequences in the Chinese educational and societal contexts. It addresses the research problem why the CET-4 washback exists in the Chinese educational context by linking washback to the validity of the interpretations and uses of the test score by the test developer and the test users.

To achieve the above research purposes, a multi-phase, multi-method research design was employed. This chapter starts with an outline of the research design (Section 3.1), illustrating how the study was implemented to answer the research questions and to achieve the research purpose. This is followed by an overview of the researcher’s role in qualitative research, particularly in interview data collection and analysis (Section 3.2). Then, some key terms in the research questions are defined in Section 3.3. Finally, the specifics of the methods used in each phase of the study are described in Section 3.4, including participants, instruments, data collection procedures, and data analysis procedures.

This research received clearance from the Queen’s University General Research Ethics Board and complies with the Tri-council Standards on research with human subjects (see Appendix A for clearance letter). In addition, prior to data collection, all participants were given a letter of information that detailed their involvement in the study and signed a consent form (see Appendix B, Appendix C, and Appendix D).
3.1 Multi-phase, Multi-method Research Design of the Study

Figure 3-1 outlines the multi-phase, multi-method research design of this study.

Phase 1

Research focus:
Test developer’s intended consequences of the CET-4 (RQ1)

Method
Qualitative document and interview data gathered and thematically analyzed

Phase 2

Research focus:
Test users’ interpretations and uses of the CET-4 in the educational and societal contexts (RQ2)

Method
Qualitative interview data gathered and thematically analyzed

Research problem:
Why does the CET-4 washback exist in the Chinese educational context

Phase 3

Research focus:
Washback of the CET-4 on students’ learning and learning outcomes (RQ3)

Method
Quantitative questionnaire and test data gathered and SEM analysis conducted

Figure 3-1. Multi-phase, multi-method research design of the study.
My choice of the multi-phase, multi-method research design in this study was determined by the purposes of the study and based on the following rationales:

First, consequences are the focus of investigation in this study. This construct is conceptualized in a broad sense from the perspective of program evaluation (see Section 1.2). It involves intended and unintended, direct and indirect consequences in the Chinese educational and societal contexts associated with the implementation of the CET-4 testing program, as well as value implications underlying interpretations and uses of the test. Investigation of this wide scope of consequences entails collecting data from multiple sources. Different kinds of consequences were investigated in three phases to address the different but interrelated research questions. Answers to these questions together addressed the research problem.

Second, consequences are linked to construct validity within an IUA framework drawing on the argument-based validation approach (Kane, 2013). Intended consequences of the test guide the IUA (Bachman & Palmer, 2010). Therefore, the test developer’s intended consequences as well as procedures to achieve these consequences were first investigated in Phase 1. Findings from this investigation constituted the baseline evidence against which findings from Phase 2 and Phase 3 are discussed. Findings derived from these two phases provide evidence on the extent to which the intended consequences are being achieved as well as evidence about unintended consequences. Based on these lines of evidence, I was able to make some evaluative arguments regarding the decision of using the CET-4 program as a policy instrument for educational reform in the Chinese context. Linking washback evidence to evidence about
construct validity and to evidence about the actual uses of the test, I was able to address the research problem.

Third, the nature of evidence to be collected in each phase of the study determined the methods used for data collection and analysis.

Phase 1 and Phase 2 collected evidence about the interpretations and uses of the CET-4. Considering that test interpretation and use are value-laden and contextually bounded (Koch & DeLuca, 2012; McNamara & Roever, 2006; Moss et al., 2006), I took an interpretive approach (Yanow & Schwarts-Shea, 2006) and used qualitative methods in these phases. This approach would provide in-depth understanding of the meaning the participants constructed (Merriam, 2009).

Phase 3 took a predominantly quantitative survey approach. Particularly, structural equation modeling (SEM) was used to investigate the washback of the CET-4. SEM is a comprehensive statistical approach for testing modelled hypotheses about relations among latent variables. Usually, a model is specified \textit{a priori} based on substantive theory or hypothesis (Raykov & Marcoulides, 2006). In this study, a structural regression model (Raykov & Marcoulides, 2006) is proposed in order to postulate the explanatory relationships among students’ perceptions of the CET-4 demands and uses, their test-taking expectations, test preparation practices, and learning outcomes. This model is based on previous washback literature (e.g., Green, 2007; Hawkey, 2006; Hughes, 1993; Shih, 2007; Xie, 2010). Using SEM, I was particularly interested in how the above factors work together, and the extent to which students’ perceptions of test demands and test uses influence their test preparation practices, which in turn influence their learning outcomes.
3.2 Researcher’s Role in This Study

In this study, I assumed the role of *independent researcher* (Cronbach, 1989), gathering and analyzing the data, and reporting the results. Phase 1 and Phase 2 of the study are predominantly qualitative in nature. Qualitative research requires the researcher’s sincere introspection to acknowledge personal biases, values, or interests (Creswell, 2003) and inclusion of some biographical details about the researcher (Denscombe, 2010). This is particularly relevant for this study given that it is related to validation, consequences, and program evaluation, all of which are value-laden. The impact of my personal experiences, identity, values, and beliefs on the research should not be entirely neglected. An independent researcher is not necessarily an objective researcher. Thus I provide the following brief account of my researcher’s self (Simons, 2009) and acknowledge my own subjectivity in the research process.

To begin with, having grown up and been educated in China, I have witnessed and experienced the tremendously important role large-scale high-stakes tests have played and are playing in Chinese society. This experience motivated me to explore the impact of such tests in China. Secondly, I was an EFL teacher myself for over 20 years. I saw the dramatic development of EFL education in China over the past few decades as well as the increasingly powerful impact of English tests in the Chinese educational context and society at large. Related to this experience are my beliefs and values about English language teaching and learning as well as my concerns over their adverse consequences, particularly on students. Thirdly, I received my academic training in educational and language assessment in North America. This experience triggered my interest in research into test consequences within a test validation framework.
There is benefit to conducting this research on the CET-4 in China, a cultural context of which I am a member. My insider role and status allowed me rapid and complete acceptance by the participants and afforded a common ground between the participants and me from which to begin the research. However, qualitative researchers (e.g., Tilley & Chambers, 1996) caution that bias and issues unique to insider research can occur that put the trustworthiness or validity of the study at risk. In this study, I dealt with two major issues: my belief about the adverse consequences of the CET-4 on students, and my assumed familiarity with the research settings. These issues may have limited my ability as a researcher to probe for deeper meaning or understanding of the phenomenon from the participants’ perspectives. In order to address these issues, self-reflection was the major strategy I used in Phase 1 and Phase 2 of this study in both collecting and analyzing the interview data. I will explain how this strategy was used in Section 3.4.3 and Section 3.4.4.

3.3 Operational Definitions of Terms in Research Questions

For operational purposes, the key terms in the three research questions in this study are defined as the following:

RQ1. What are the intended consequences of the CET-4 and what are the major procedures taken by the test developer to achieve the intended consequences?

*Test developer of the CET-4* includes both the CMoE and the National College English Testing Committee (NCETC).

*Intended consequences* are understood from two interrelated perspectives. From the NCETC’s perspective, intended consequences focus on the utility of the CET-4 score for its legitimate uses in the educational context. From the CMoE’s perspective, *intended*
consequences refer to the intended beneficial effects of implementing the CET-4 in the Chinese tertiary educational context. Based on previous literature, these consequences may include promoting motivation, providing feedback, and enhancing the alignment between curriculum, assessment, and teaching and learning.

Major procedures by the test developer refer to both the measures taken by the NCETC at the technical level to enhance the quality of the CET-4 and the measures at the policy-level by the CMoE to support the decision rule.

RQ2. What are the actual uses of the CET-4 in the educational, business, and government contexts and what are the value implications underlying the actual uses?

Actual uses of the CET-4 refer to the uses of the CET-4 based on the test users’ interpretations of the information provided by the test score. Test users refer to those engaged in making decisions involving the uses of CET-4 scores. The educational, business, and government contexts refer to the educational institutions, business companies, and government institutions where the CET-4 score was used. Value implications refer to the values shaping the test users’ interpretations and uses of the CET-4 score.

RQ3. How does the CET-4 influence students’ learning and learning outcomes?

Students refer to those preparing for the CET-4. Learning refers to what students report they did in preparation for the CET-4. Learning outcomes refer to students’ performance on the CET-4 test administered for this study.

3.4 Methods

This section presents the specifics of the methods used in this study including participants, instruments, data collection procedures, and data analysis procedures.
3.4.1 Participants

Participants in this study include a participant from the NCETC, the CET-4 test developer; eight CET-4 test users from educational and societal contexts; and 416 undergraduate students from two Chinese universities. The sample size was determined based on the methods used for data analysis in this study (see Section 3.4.4), as well as the feasibility of data collection, considering the marking procedures of the test paper (see Section 3.4.2) and the time constraint of this research.

Participant from the NCETC

Deyi² from the NCETC was invited to participate in this study. She holds a PhD in applied linguistics and has been involved in language assessment research and teaching for over 20 years. She has been working on the NCETC since 1999. Therefore, she could provide rich information about the test developer’s interpretation of the CET-4 score and the intended uses of the score. The current NCETC was set up in 2004. It consisted of 25 professors from 23 universities across China, mostly heads of departments of English in their universities. The NCETC is responsible for the design, development, and administration of the CET, as well as providing advice on CET-related policies. The NCETC members meet on a regular basis for item writing and revision, standard-setting, essay marking, training of markers, score equating and reporting, and test revision.

Student participants

Four hundred and sixteen second year undergraduate students in two universities participated in this study. They were non-English majors preparing for the CET-4 in

---

² All the names of the participants in this thesis are pseudonyms to protect their identities.
December of 2011. The students were recruited using cluster sampling from 11 classes (The class size of the College English course in Chinese universities is approximately 50) taught by 6 different teachers in the key university in City A in the south of China (n=212) and the non-key university (n=204) in City B in the north of China. This sampling was based on the following considerations. First, key and non-key universities in China represented different instructional contexts in terms of factors such as teaching and learning resources. Second, graduates from these two types of universities often had different opportunities of employment, thus may have different perceptions of the uses of the CET-4 score assuming the CET score was used for that purpose. Third, second year undergraduate students were chosen because students usually take the CET-4 at the end of the first term in their second undergraduate year.

Of the 416 student participants, 51.2% were male; 48.8% were female. Their major areas of study included finance, accounting, commerce, biology, mechanical engineering, economics, computer science, and applied chemistry. On average, they had learned English for 9.42 years at the time of the study.

**CET-4 test user participants**

Eight CET-4 users from the educational and societal contexts participated in this study. Table 3.1 summarizes information about these participants.
Table 3.1

Profile of Test User Participants

<table>
<thead>
<tr>
<th>Context*</th>
<th>Name*</th>
<th>Position</th>
<th>Affiliation</th>
<th>Description of the affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>Gongli</td>
<td>Director</td>
<td>Employment Promotion Office affiliated with the Municipal Human Resource and Social Development Bureau of a metropolis with highly developed economy in China (City A)</td>
<td>A municipal office responsible for promoting employment of young people, particularly university graduates; affiliated to the Municipal Human Resource and Social Security Bureau</td>
</tr>
<tr>
<td></td>
<td>Gongfei</td>
<td>Deputy director</td>
<td>Municipal Human Resource and Social Development Bureau of a medium-sized city with medium-level economic development in China (City B).</td>
<td>A component department of the Municipal Government of City B; its major responsibilities included: implementing laws, regulations, rules, and policies concerning human resources and social development; researching, drafting, and implementing local regulations, rules, policies, and plans concerning human resources and social development.</td>
</tr>
<tr>
<td>Education</td>
<td>Edi</td>
<td>Associate dean of College English Department</td>
<td>A university in City A</td>
<td>The university was a key university, with approximately 8,000 undergraduate students. Each year, about 3,500 students took the CET.</td>
</tr>
<tr>
<td></td>
<td>Edu</td>
<td>Dean of College English Department</td>
<td>A university in City B</td>
<td>The university was a non-key university, with approximately 13,000 undergraduate students. Each year, about 7,000 students took the CET.</td>
</tr>
<tr>
<td>Context*</td>
<td>Name*</td>
<td>Position</td>
<td>Affiliation</td>
<td>Description of the affiliation</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>---------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Business English-relevant</td>
<td>Buran</td>
<td>Human resource manager</td>
<td>A branch in City A of a Singaporean property management company</td>
<td>Established in 2005 with 132 employees; over half of its clients were international; English proficiency was required for most positions; 3-5 university graduates were recruited each year.</td>
</tr>
<tr>
<td></td>
<td>Buru</td>
<td>Human resource manager</td>
<td>A joint-venture electronics company in City B</td>
<td>Established in 1999, with 130 employees; recruited about 10 university students each year. English was required to read and write emails, documents and technical reports in the workplace.</td>
</tr>
<tr>
<td>Not English-related</td>
<td>Buna</td>
<td>Human resource manager</td>
<td>A state-owned pharmaceutical company in City B</td>
<td>Established in 1966, with 7,800 employees; recruited approximately 30 university students each year. Most of the positions were not English-related.</td>
</tr>
<tr>
<td></td>
<td>Buni</td>
<td>Human resource manager</td>
<td>A subsidiary in City A of a state-owned national bank</td>
<td>Provides a full range of financial products and services to personal and business customers; the majority of the customers were Chinese; with over 8,000 employees; approximately 150 university graduates were recruited each year. Most of the positions were not English-related.</td>
</tr>
</tbody>
</table>

*All the names of the participants in this thesis are pseudonyms to protect their identities. The initial letters of each pseudonym indicate the contexts where the participant was recruited. Go: government; Ed: education; Bur: Business where English was relevant to the workplace; Bun: Business where English was not relevant to the workplace.*
These test user participants included 1) two heads of English departments of the two universities where the student participants were recruited; 2) two government officials from the departments of human resource and social development of the two cities where the two universities were situated; 3) two human resource managers of two companies in each of these two cities where the government official participants were recruited. Maximal variation sampling (Creswell, 2008) was used in recruiting these participants. In recruiting the city government officials, the major dimension considered was the level of the social and economic development of the city. In recruiting participants from universities, the major dimension considered was the prestige of the university. In recruiting participants from the business context, the major dimension considered was the relevance of English in the company. Details about the process of participants recruitment is presented in Section 3.4.3.

3.4.2 Instruments

In Phase 1 and Phase 2 of the study, semi-structured interviews were conducted with the participant from the NCETC and the eight test user participants in the educational and societal contexts. The rationale for semi-structured interviews is that they afford flexibility to the participants’ responses (Creswell, 2008). The purpose of using semi-structured interviews is to capture the informants’ accounts of their interpretations and uses of the CET-4 score.

The guideline questions for the semi-structured interview with the participant from the NCETC (Appendix E) were developed based on the CET-4 IUA framework illustrated in Figure 2-2. These questions focused on the explanation of the CET-4 score and the intended uses of the test. Questions concerning the inferences involved in the
IUA, including warrants and assumptions underlying them, were included. There were also questions about how the assumptions were backed up by evidence from the perspective of the test developer.

The guideline questions for the semi-structured interviews with the CET-4 test users (Appendix F) focused on how the CET-4 scores were used in different contexts, the test users’ interpretations of the test scores, and the values the participants held in interpreting and using the test results. There were also questions seeking demographic information about the participants and contextual information about the organizations.

To collect the test score data, I constructed a test paper by combining different sections of authentic CET-4 papers administered since 2006. The content, item types, score value, and time length for each part of the CET-4 are presented in Table.1.1. Sample items on this paper are presented in Appendix G. The constructed test paper was reviewed by the NCETC participant for accuracy and validity. Keys to the questions on the test paper were provided by the NCETC participant as well.

Alongside the test, a questionnaire was used to collect data about students’ perceptions of the CET-4 test demands and test uses, their CET-4 test-taking expectations as well as their test preparation practices. The questionnaire also collected demographic data about the students. The questionnaire was developed based on previous research (Green, 2007; Hawkey, 2006; Xie, 2010) and results from Phase 1 and Phase 2 of the study also informed its development. The questionnaire consisted of five major sections. The first section included 44 items that measured students’ perceptions of the CET-4 demands (i.e., their perceptions of the knowledge, skills and strategies necessary for answering the CET-4). The second section included 16 items measuring students’
perceptions of the CET-4 uses (i.e., their perceptions of the usefulness and values of the CET-4 score). Items in these two sections used a 7-point visual analog scale, with 1=strongly disagree; 7=strongly agree. The third section of the questionnaire included 11 items that gathered data concerning students’ self-efficacy in English proficiency in relation to the test demands (again, using a 7-point visual analog scale) and 10 items that measured their perceptions of the CET-4 task difficulty using a 7-point scale, with 1=very easy; 7=very difficult. The fourth section included 39 items investigating students’ CET-4 preparation practices (i.e., the extent to which they used different types of learning practices, such as cramming, memorizing, rehearsal of test-taking skills, and use of long-term, skill-development strategies). A 7-point visual analog scale was used in this section, with 1=never; 7=always. The last section of the questionnaire gathered demographic data of the students such as gender, educational background, experience of English language learning, and their familiarity with the CET-4.

Table 3.2 illustrates the scales and sub scales measured by the questionnaire.

Table 3.2.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Sub construct</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptions of test demands</td>
<td>Listening</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Integrated Skills</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Writing &amp; Translation</td>
<td>9</td>
</tr>
<tr>
<td>Perceptions of test uses</td>
<td>Instrumental uses</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Achievement uses</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Task values</td>
<td>4</td>
</tr>
<tr>
<td>Test-taking expectations</td>
<td>Self-efficacy</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Perceptions of task difficulty</td>
<td>10</td>
</tr>
<tr>
<td>Test preparation practices</td>
<td>Test preparation management</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Cramming target knowledge &amp; skills</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Rehearsal test-taking strategies</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Long-term skill development</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Social-affective strategies</td>
<td>8</td>
</tr>
</tbody>
</table>
The questionnaire was written in Chinese to ensure the participants’ accurate understanding of the items (see Appendix H for the English translation of the questionnaire).

3.4.3 Data Collection

In Phase 1 of the study, I collected policy documents disseminated by the CMoE and the NCETC concerning the implementation of the CET-4. These documents included the College English curriculum, the CET-4 Syllabus, minutes of press release, and documents posted on the NCETC official website. In total, I collected 17 documents. I thematically analyzed these documents based on the IUA framework presented in Figure 2-2 and developed interview questions with Deyi, the participant from the NCETC.

I contacted Deyi, explained to her my research plan and sent her the Letter of Information and the Consent Form. At the request of Deyi, I also sent her some literature I cited (e.g., Kane, 2006; Chapelle et al., 2010) as well as the IUA framework for the CET-4 I constructed. One month later, I conducted a 63-minute in-depth interview with Deyi at her office. The interview was conducted in Chinese and was digitally recorded. I took brief notes during the interview.

Both Deyi and I were interested in language assessment research and had experiences of teaching EFL in Chinese universities. This commonality provided a common ground from which to begin the interview. However, the commonality might have also led to an emphasis in the interview on shared factors between Deyi and me and a de-emphasis on factors that were discrepant. In order to address this issue, when Deyi
implied during the interview that I knew what she was talking about, I would delve deeper and ask for clarification and elaboration. Also, I used the IUA for the CET-4 to shape and guide the interview to avoid the undue influence of the shared factors between Deyi and me.

In Phase 2, I contacted the graduate employment offices of the two universities where the student participants were recruited and obtained two lists of companies hiring undergraduate students. To sample test user participants from the lists, one dimension I considered was the use of English in the workplace. English may or may not be job-related in the workplace. Employers from these different settings who required CET-4 scores may present different perspectives on the interpretations and uses of the test scores. Based on this consideration, I screened the lists for companies that required graduates’ CET-4 scores. I chose six companies from the two lists based on their business features as well as the descriptions of job responsibilities on their job advertisements: three companies where English was job-related and three companies where English was not job-related. These six companies were contacted with the assistance of the staff from the graduate employment offices. Two of them declined the invitation for an interview. I asked participants from the other four businesses to sign the consent forms. Next, I conducted a telephone interview with Buran and individual face-to-face interviews with all the other participants in their offices. Each of the interviews was around 45 minutes. The interviews were digitally recorded. I took brief notes over the interviews. The interviews were conducted in Chinese.

In addition to the test users in the business context, I also recruited two government official participants from the two cities mentioned above. The two officials
were contacted through my friends in these cities. I sent them the letter of information and the consent form. Upon obtaining consent from them, I conducted an individual face-to-face interview with each of them. The interview with Gongfei was conducted in his office and the interview with Gongli was conducted at a quiet location of his choice. Both of the interviews were conducted in Chinese and digitally recorded. The interview with Gongli was 45 minutes and the interview with Gongfei was 40 minutes. I took brief notes over both the interviews.

My major concern about the interviews with the above test user participants was the depth of the data collected. I noticed that the participants generally held a defensive attitude when I contacted them for the interviews and did not explain their individual perspectives fully in the interviews. To address this issue, in correspondence requesting participation, I assured the participants of the information confidentiality and participant anonymity. At the beginning of the interviews, I reaffirmed to them my role as a researcher and clarified with them how the data would be used and shared.

When I contacted the two test user participants in the educational context (Edi and Edu), they accepted my invitation for participation immediately. I conducted individual face-to-face interviews with them. Each of the interviews was about 45 minutes and was conducted in Chinese in the participants’ offices. Both of the interviews were digitally recorded and I also took brief notes over the interviews. Edi and Edu were open to my questions and looked comfortable sharing their dilemmas and frustrations over the interviews. However, I noticed that sometimes they would make assumptions about my familiarity and understanding. To address this issue, I would ask them to clarify their points in relation to their local contexts.
In Phase 3 of the study, I collected test score data and questionnaire data from the student participants. To ensure the validity and reliability of the questionnaire measurement, the following procedures were followed before the data collection.

First, a pretest of the questionnaire was conducted with 10 of my students\(^3\). They were invited to read and respond to the questions, and then were asked about the clarity of the items on the questionnaire. Based on their comments, the questionnaire was revised. Second, two experienced College English teachers were invited to review the revised questionnaire for validity. Further revisions were made based on their feedback. Third, the revised questionnaire was pilot tested using an iterative process. Forty second-year undergraduates from one key university (\(n=20\)) and one non-key university (\(n=20\)) were invited to answer the questionnaire. Cronbach alpha internal consistency coefficients for the four sections of the questionnaire ranged from .79 to .91. Revisions were made based on the pilot test results. Then, the revised questionnaire was issued to another 40 second-year undergraduates from one key university (\(n=20\)) and one non-key university (\(n=20\)) before it was finalized.

Five hundred and fifty-eight second year undergraduate students from the two universities were invited through their College English classroom teachers to participate in this study. Of these students, 499 accepted the invitation and signed the consent form. The questionnaire was issued to these students in their classrooms in mid-November 2011 by their classroom teachers. By the time of the data collection, the participants had studied College English for one year and were preparing for the CET-4. It took the

\(^3\) I was teaching two classes of College English while I was collecting data for this study. However, the participants in this study did not include any students from my classes.
participants about thirty minutes to complete the questionnaire. The completed questionnaires were collected by the classroom teachers and then given to me.

Two weeks before the participants took the official CET-4 (the second week of December, 2011), the CET-4 test paper was administered in their classrooms by the researcher with the assistance of the classroom teachers as a practice test. Of the 499 participants who completed the questionnaire, 416 took the test. Responses to the multiple choice test questions were scored by a scoring machine and the writing section was scored by two experienced CET-4 raters recommended by Deyi using the official scoring rubrics. As the primary researcher, I resolved disagreements between the two raters over three points. The scores assigned by the raters were averaged and the average score was used for data analysis. I scored the other constructed response questions in the test such as the Compound Dictation, the sentence completion items in Skimming and Scanning as well as the Translation items, using the keys provided by Deyi.

3.4.4 Data Analysis

The IUA for the CET-4 presented in Figure 2-2 was used as an analytical framework both to analyze data concerning the test developer’s intended consequences and efforts to achieve the consequences and to synthesize findings of this study.

In order to answer RQ1, first, official documents disseminated by the CMoE and the NCETC were analyzed deductively. Content analysis was conducted to organize information into categories related to the claims, inferences and their underlying assumptions in the IUA. Second, the transcript of the interview with Deyi was analyzed to deduce themes relevant to the test developer’s interpretation and intended uses of the CET-4 score as well as the efforts to support the IUA. The interview data were analyzed
inductively using King and Horrocks’ (2010) thematic analysis system (see Figure 3-2). Thematic analysis is considered most appropriate for analyzing interview data (King & Horrocks, 2010). Inductive thematic analysis was used to analyze the interview data in this study because it allowed me to 1) draw interpretations consistent with the data collected; 2) associate an analysis of the frequency of themes and codes with the targeted research questions; and 3) apply or link codes to raw data for later analysis (e.g., comparing the relative frequencies of codes across the different interview datasets). In Figure 3-2, the coding process is broken down into a series of stages (and steps within them). However, my analysis of the interview data did not always progress in a purely sequential manner. I often cycled back and forth between these stages in my data analysis.

Figure 3-2. A thematic analysis system. Adapted from King & Horrocks (2010, p. 153).
Figure 3.3 demonstrates some of the codes from the transcript of the interview with Deyi and how these codes were grouped into interpretive codes, which in turn were aggregated together to form an overarching theme.

**Figure 3.3.** Demonstration of three coding levels.

In order to answer RQ2, immediately after each interview with a participant, the interview data were analyzed inductively using King and Horrocks’ (2010) thematic analysis system. The interview with the second participant in the same context was then conducted and the data were treated in the same way. After that, the two interviews were compared. Fragments from the interviews that had been given the same codes were compared. If one code was mentioned by one participant but not the other in the same context, the code was also included in the reported results to illustrate the perspectives of the test users in the context. The coding process identified a large number of overlapping codes associated with the context. I grouped these interview transcripts into four datasets based on the contexts: the educational context; the government context; the English-related business context; and the English-unrelated business context.
All interviews, including the eight interviews with the test user participants and the interview with the test developer participant, were fully transcribed. Only the parts of the transcripts that are quoted in this dissertation were translated into English. The extended process of transcription gave me a chance to get closer to and familiar with the qualitative data. Considering that the interview dataset was not very large (81 pages of transcripts), qualitative computer software programs were not used.

In analyzing the interview data, the dual role of an insider and a researcher resulted in my role confusion. Sometimes I found it difficult separating my own perceptions, which might be clouded by my personal experiences, from perceptions and perspectives of the participants. To address this issue, I first identified my own thoughts, beliefs, and assumptions before the data analysis process, wrote them down and put them aside for reflection during the data analysis process. Second, I used “member checks” (Lincoln & Guba, 1985). I asked the participants to make clarifications and to check the accuracy of my analytic categories, interpretations, and conclusions.

In order to answer RQ3, the questionnaire data and the test score data collected in Phase 3 were analyzed quantitatively. SPSS 17.0 and LISREL 8.70 were used for the data analysis. First, all the data were entered into the SPSS program manually. Then, missing value analysis was conducted. Both the questionnaire and the test score datasets had less than 5% missing values and the missing pattern did not have an obvious pattern. Tabachnick and Fidell (2005) state “if only a few data points, say 5% or less, are missing in a random pattern from a large data set, the problems are less serious and almost any procedure for handling missing values yields similar results” (p. 63). In this study, the mean substitute method was used for missing values. Outliers were examined at the item
level through SPSS stemleaf and boxplot. A few outliers associated with data entry errors were identified and corrected by revisiting the original data. After outlier transformation, I checked all items to determine if their distribution met the univariate normality and multivariate normality assumption through examining skewness, kurtosis statistics, and the graphics (i.e., histogram, and expected normal probability plots). Multivariate normality was checked using Prelis with LISREL.

Descriptive statistics were used to summarize: 1) students’ perceptions of the CET-4 test demands; 2) their perceptions of the test uses; 3) their expectations of CET-4 test taking; and 4) their CET-4 preparation practices. Principal axis factoring (PAF) was conducted for these scales in the questionnaire. The Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy for these variables were .97, .93, .89, and .96 respectively. Bartlett’s tests of sphericity were significant (p<.001) for all these variables. These statistics indicate that it was appropriate to use PAF. Varimax was used as the method for rotation in the PAFs to minimize the number of variables with high loadings on a factor. To guide components selection, two standards were set: 1) they obtained eigenvalues of 1 or above, and 2) they were located at or above a point of inflexion (or ‘elbow’) on a scree plot. Composite factor scores were calculated for each case by taking the mean of items that loaded on a given scale. For easy interpretation, composite factor scores were used in subsequent analyses to represent the values of the construct indicators in the following SEM analysis.

In this study the following washback model (Figure 3-4) was proposed based on previous research into washback on student learning (e.g., Green, 2007; Hughes, 1993; Hawkey, 2006; Shih, 2007; Xie, 2010).
The model outlined by Figure 3-4 assumes that students’ perceptions of the CET-4 test demands and uses as well as their test-taking expectations influence their test preparation practices, which in turn influence their learning outcomes. These hypothesized relationships among the latent variables were assessed against empirical data collected to see if they hold and path coefficients among the latent variables were estimated to examine the relative strength of the relationships. To evaluate model fitness, the following fit indices were used: Chi-square value; standardized root mean square residual (SRMR; Bentler, 1995); comparative fit index (CFI; Bentler, 1989, 1990), root-mean-square error of approximation (RMSEA, Steiger & Lind, 1980); and adjusted goodness-of-fit index (AGFI, Jöreskog & Sörbom, 1986). The following criteria were used to determine model fit based on suggestions by Hu and Bentler (1999): SRMR ≤ .08, CFI ≥ .95, AGFI ≥ .90, and RMSEA ≤ .06.

Results from the qualitative analysis in Phase 1 and Phase 2 of the study are reported in Chapter 4 and results from the quantitative analysis in Phase 3 are reported in Chapter 5. These results are synthesized and discussed in Chapter 6 within the IUA framework.
Chapter 4 Intended Consequences and Actual Uses of the CET-4: Test Developer’s and Test Users’ Perspectives

This chapter reports findings regarding 1) intended consequences of the CET-4 as described by the test developer; 2) major procedures taken by the test developer to realize the intended consequences; 3) actual uses of the CET-4 results in the education, business, and government contexts; and 4) value implications underlying actual uses of the test from the perspectives of the test users. These results address RQ1 and RQ2 of this study:

RQ1: What are the intended consequences of the CET-4 and what are the major procedures taken by the test developer to achieve the intended consequences?

RQ2: What are the actual uses of the CET-4 in the educational, business and government contexts and what are the value implications underlying the actual uses?

In order to address RQ1, three sources of data were analyzed: interview data with the test developer participant (Deyi); documents disseminated by the NCETC and the CMoE; and documents from the official website of the NCETC. The document data were analyzed deductively using content analysis (Hsieh & Shannon, 2005). The IUA framework (see Figure 3-2) was used for the deduction. The interview data were analyzed inductively using thematic analysis (King & Horrocks, 2010).

To answer RQ2, the interview data with the test user participants in the education, business, and government contexts were analyzed using inductive thematic analysis. Detailed information about the participants is provided in Section 3.4.1.

Results reported in this chapter focus on categories generated from qualitative analyses of both the document and interview data mentioned above. Procedures used to generate these categories are presented in Section 3.4.4. This chapter consists of five
sections. Section 4.1 and Section 4.2 report categories regarding the CET-4 developer’s intended consequences of the test and the major procedures to achieve these consequences. These results answer RQ1. Section 4.3 and Section 4.4 present categories related to the actual uses of the CET-4 in the education, business, and government contexts, and value implications underlying these uses. These results answer RQ2. Section 4.5 ends this chapter with a brief summary of the major findings.

4.1 Intended Consequences of the CET-4

My understanding of the intended consequences of the CET-4 was based on analysis of the documents disseminated by the test developer including the NCETC and the CMoE, and my interview with Deyi. The NCETC was responsible for test development, administration, scoring, and reporting the test score, particularly at the technical level. The CMoE initiated the CET-4 and provides guidelines for the test reform at the policy level. It also appoints the NCETC members.

Analysis of the document and interview data generated two interrelated categories regarding the intended consequences of the CET-4: 1) to provide useful information about students’ English language proficiency based on the curriculum; and 2) to promote College English education in China. Table 4.1 summarizes these categories with their related codes as well as the sources and frequencies.
Table 4.1

Intended Consequences of the CET-4

<table>
<thead>
<tr>
<th>Findings</th>
<th>Frequency by source of data</th>
<th>Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interview</td>
<td>Document</td>
</tr>
<tr>
<td>Category 1: The CET-4 is intended to provide useful information about students’ English proficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Codes:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-is scientific, objective, accurate, and fair measure of students’ English language proficiency</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>-is relevant to College English teaching and learning</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>-is interpretable to users</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Category 2: The CET-4 is intended to promote College English education in China.

| Codes:                                                                 |            |          |         |               |
| -is initiated to implement EFL policy and to motivate EFL learning in Chinese colleges and universities | 1          | 6        | 3        | 10            |
| -is intended to influence EFL teaching and learning in Chinese colleges and universities | 5          | 7        | 4        | 16            |
| -is intended to inform EFL teaching and learning in Chinese colleges and universities | 7          | 12       | 5        | 24            |

4.1.1 To Provide Useful Information about Students’ English Proficiency

A major intended consequence of the CET-4 was that the test score would provide users in the educational context with useful information about students’ English proficiencies based on the College English curriculum (i.e., the CECR). The usefulness of the information provided by the CET-4 score was based on the assumption about the quality of the test. Analysis of the documents showed that *scientific, objective, accurate,*
and fair were terms most frequently used by the test developer to describe the quality of the CET-4. These terms were highlighted on the home page of the NCETC official website (http://www.cet.edu.cn/). The first CET-4 test syllabus (CET Design Group, 1987) stated that the test was intended to provide an objective, scientific, and fair measure of the Chinese undergraduates’ English proficiency. In the most recent curriculum, the CECR, the CET-4 is considered a significant component of a “comprehensive, objective, scientific, and accurate assessment system.” The excerpt below is my translation from the CECR:

Assessment is a key component of College English curriculum. A comprehensive, objective, scientific, and accurate assessment system is of vital importance to the achievement of the objectives in the curriculum. It provides important evidence for teachers to obtain feedback information, improve teaching and management, and ensure teaching quality. It also provides students with an effective means to adjust their learning strategies and methods, improve their learning efficiency, and achieve the desired learning outcomes (CMoE, 2007a).

In the interview, Deyi explained the usefulness of the CET-4 by highlighting the relevance of the test to College English teaching and learning. She repeatedly emphasized that the CET-4 was an educational assessment in nature, intended to serve educational purposes. Deyi particularly highlighted the relevance of the recent reform of the CET-4 score reporting system to College English education. She said that the major consideration underlying reporting the sub-scores is how the test can provide feedback to teaching and learning.
In order to illustrate how the CET-4 could inform English language learning, Deyi gave me the following example in the interview:

I knew a student who had been recommended for doctoral studies, but was not able to pass the CET cut-off score required by the university. So his father came to see me, saying that he’s a very promising student, but due to his low CET score, he was not able to make it. I don’t want to make any comments on the use of the test score here. This is out of the control of the test developer. I asked him [the student] to bring his CET score report for me to take a look and found that his reading score was very high, but his writing and listening scores, in contrast, were very low. Based on his CET sub-scores, I talked with him about what he could do to improve his writing and listening. So reporting the CET score in this way enables students to have a better understanding of their strengths and weaknesses in English proficiency.

In order for the test users in the educational context to effectively use the information provided by the CET-4 score, validity of their score interpretation is a key issue. With regard to this issue, Deyi expressed her concerns over teachers’ and educational administrators’ “misinterpretations and misuses of the percentile score.” She pointed out that the CET-4 score was not being interpreted by test users in the way that was intended by the test developer due to their lack of assessment literacy and the social context that shaped their interpretations:

It is very hard for people to interpret the percentile score. For example, one teacher once asked me: “Could you tell me how I could convert the CET-4 score to a percentage score?” You know, in some universities and colleges, they simply
divide the students’ scores by 7 to convert them to percentage scores because they think that the full score used to be 100, and now it is 700-ish, so by dividing the score by 7 it can be converted to a percentage. This is an issue related to assessment literacy. Test users need to have a good understanding of standardized test. This is particularly true in the Chinese society. Test scores have been traditionally reported in the form of percentage in China. In this context, test scores reported in other forms are more likely to be misinterpreted.

In order to enhance the interpretability of the norm-referenced score, the NCETC provides detailed explanations of the concept of percentile score in the test syllabus and on its official website. The formula for converting the raw score to the reported standardized score is provided with some examples. A table is also provided whereby test users can identify a student’s percentile position based on the reported score. Appendix I is an excerpt from the CET website concerning the test score interpretation.

In summary, utility of the CET-4 for College English teaching and learning is a major intended consequence of the test. The utility of the CET-4 to its users in the educational context is based on three assumptions. First, the test is claimed to be a scientific, objective, accurate, and fair measure of students’ English proficiency. Second, the reported CET-4 score is intended to provide information relevant to College English teaching and learning. Third, the test score should be interpretable to test users and they should interpret it in the way that is intended by the test developer.

4.1.2 To Promote College English Education in China

The CET-4 was initiated and is being implemented as a policy instrument together with the College English curriculum for the purpose of promoting EFL education in
Chinese universities and colleges (i.e., College English education). This purpose is consistently stated across the documents disseminated by the CET-4 developer. To achieve this goal, the curriculum specifies the expected standards and suggests approaches to teaching. The CET-4 testing program is implemented to measure students’ achievement of these standards.

Over the past few decades, the goal of College English education has undergone significant changes with the changing Chinese educational and societal contexts. The College English curriculum and the CET-4 have been revised accordingly. In the 1980s shortly after China opened its door to the world, EFL education began to receive policy attention in the country. At that time, reading was considered the primary means of communication with the English-speaking world. To implement the EFL policy, the first College English curriculum (CMoE, 1985, 1986) was issued, which gave priority to English reading ability in defining English proficiency. Meanwhile, the CET-4 was initiated with the intention to motivate English teaching and learning in higher education sectors (CET Design Group, 1987). All non-English majors were required to pass the test to graduate in the Chinese colleges and universities. In the original form of the CET-4, reading was given the highest weighting, accounting for 40% of the total score.

The turn of the 21st Century witnessed China’s further “opening-up.” Particularly, its accession to the World Trade Organization in 2001 and its successful bid for the 2008 Olympic Games contributed to a rise in the prominent status of EFL. Quality EFL education was considered an essential component of quality higher education, which in turn was considered an important contributing force to the nation’s strength in the global knowledge economy (CMoE, 2006a, 2007a). Meanwhile, China’s tertiary education
expanded dramatically. In 1995, only 5% in the age group of 18-22 had access to higher education; in 2007, the ratio increased to 23% (CMoE, 2007b, 2008). Annual new tertiary student enrolment did not reach one million until 1997; a decade later it exceeded 5.6 million. There was an increase of about half a million each year from 1999 to 2006. Such an exponential increase generated great concerns over the quality of China’s tertiary education (CMoE, 2002a). Particularly, College English education was criticized for its “low efficiency and effectiveness” in improving students’ English use ability (Dai, 2001; Hu, 2005; Jing, 1999). Consequently, there were demands from different stakeholders such as students, parents, and employers for quality EFL education to increase students’ English communicative abilities.

In this context, the College English curriculum was updated in 1999 (CMoE, 1999) for the purposes of: 1) addressing the issue of “low efficiency and effectiveness”; 2) promoting the use of communicative language teaching approach in the classroom; and 3) better serving the growing requirement in the society for English communicative abilities of university graduates. Following this update, a variety of constructed response questions were introduced in the CET-4, and the weighting of the MCQs was reduced from 85% to 75%. More prominence was given to writing. Certain points would be deducted from the total score if a test-taker failed the writing part, and a zero in the writing part would fail the test taker on the whole test regardless of his/her performance on the other parts of the test. These changes were made to encourage communicative teaching and to discourage teaching to the test in the classroom.

More recently, the Chinese government launched the Higher Education Quality and Reform Project (HEQRP) to improve the overall quality and achieve sustainable
development of its higher education. The College English Reform Program (CERP) was launched as an important part of the HEQRP. The CERP involves two major components: promulgating the updated curriculum (i.e., the CECR) and reforming the CET-4. The CECR emphasizes development of students’ English use abilities, especially in listening and speaking. Promoting SRL is also considered a goal of College English education. To achieve this goal, the CECR proposes an assessment model which emphasizes the importance of assessment in informing teaching and learning.

To accommodate these changes in the goals of College English education, new task types were introduced in the CET-4 such as comprehension of longer conversations, fast reading (i.e., skimming and scanning), and Chinese-to-English translation. The component of vocabulary and structure assessed in the format of MCQ was replaced by the task of banked cloze (i.e., understanding the contextual meaning of vocabulary in a reading passage). In the current version of the CET-4, constructed response items contribute to 30 to 45% of the total score. The weighting of listening was increased from 20% to 35%. To provide more relevant information for test users, four sub-scores were reported in addition to the total score: listening comprehension, reading comprehension, cloze or error correction, and writing and translation.

A consistent purpose of the CET-4 is to promote College English education, which is value-laden in the Chinese political, social, and educational contexts. This purpose highlights the social dimension of the test as policy implementation. However, there has been a shift over the past few decades of the function the test is intended to play.

---

4 While I was writing this dissertation, the CET-4 score reporting system was revised again in December 2013. Now the CET-4 reports the total score and three sub-scores (Listening—35%, Reading—35%, and Writing and Translation—30%).
in the Chinese tertiary education context, from motivating students to study English to influencing and informing teaching and learning.

4.2 Procedures to Address Consequences of the CET-4

Categories related to the major procedures taken by the CET-4 developer to address consequences of the test are presented in Table 4.2, with their related codes as well as sources and frequencies.

Table 4.2

Matrix of Findings on Procedures to Achieve Intended Consequences

<table>
<thead>
<tr>
<th>Finding</th>
<th>Frequency by Source of data</th>
<th>Total frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1: Procedures are taken to ensure quality of the CET-4. Codes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-trains markers and monitors marking for reliability</td>
<td>5, 3, 0</td>
<td>8</td>
</tr>
<tr>
<td>-standardizes test procedures for reliability</td>
<td>5, 4, 2</td>
<td>11</td>
</tr>
<tr>
<td>-aligns with the curriculum for validity</td>
<td>6, 5, 3</td>
<td>14</td>
</tr>
<tr>
<td>Category 2: Procedures are taken to increase positive consequences and decrease negative consequences. Codes:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-revises test content and format</td>
<td>4, 8, 6</td>
<td>18</td>
</tr>
<tr>
<td>-proposes an assessment framework</td>
<td>2, 9, 3</td>
<td>14</td>
</tr>
<tr>
<td>-reforms score reporting</td>
<td>3, 8, 5</td>
<td>16</td>
</tr>
</tbody>
</table>

The first category primarily concerns procedures by the NCETC to ensure the quality of the CET-4. The second category concerns procedures both at the test design
level and at the policy level to increase the intended positive consequences and decrease
the unintended negative consequences.

4.2.1 Procedures Are Taken to Ensure Quality of the CET-4

The test developer stated that the CET-4 is a large-scale test in strict compliance
with quality requirements for standardized testing, with a complete system in place for
test development, review, administration, statistical analysis, and score reporting
(www.cet.edu.cn). In the interview, Deyi explained in much detail how considerations of
test quality directed the CET-4 developer’s efforts at all stages of test development,
scoring, and analysis. She emphasized that reliability and validity are particularly critical
in the CET-4 context, considering its large scale as well as its impact in society.

To Deyi, reliability is closely related to fairness in the sense of equality (i.e.,
students are treated equally throughout the testing process). She highlighted three factors
explaining why reliability and fairness are crucial issues in the CET-4 marking process:
the large scale of the test, time constraints, and the increase in the number of subjective
types of test questions:

You know, for each test administration, there are over 9 million test-takers.
Moreover, we have purposefully increased the number of the subjective types of
test questions. In the 1980s when the CET was started, objective types of
questions accounted for 85% of the test. At that time, you just needed to focus on
marking a composition. Things are totally different now. There are so many
subjective types of questions, and this trend will continue. It takes the 12 marking
centers around two weeks to mark all the subjective types of questions. The
number of markers in each marking center range from 200 to 400. Reliability in the marking process is crucial to ensure the quality of the test.

Deyi’s concerns over the reliability of marking the large number of constructed response questions revealed the CET-4 developer’s dilemma. On the one hand, there is the tendency to increase the percentage of constructed response questions in the test to induce its intended consequences and minimize its unintended consequences on teaching and learning. On the other hand, this tendency, coupled with the large test taker population, presents challenges to the CET-4 developer in terms of marking. According to Deyi, marker training and marking monitoring are the major quality control procedures to ensure the reliability of marking. In the interview she explained in much detail how the markers are trained and how the marking process is monitored. For example, she said:

We have an online marking program where we have integrated a marking quality control system including marker training, sampling, checking, etc. Both marker training and marking are centralized using Local Area Network (LAN) at different locations…. We use many parameters to monitor the marking quality. We have a real-time monitoring system to monitor the marking quality of each individual marker…. Reliability for CET involves intra-marker reliability, inter-marker reliability, intra-center reliability, and inter-center reliability.

In addition to quality control in the marking process, Deyi also highlighted standardization as another measure to ensure the reliability and fairness of the CET-4. She said, “To ensure consistency, the whole test administration process is completely standardized…. For such a large-scale test, without standardization, quality control would be impossible.”
Validity is another major consideration of the CET-4 developer. Deyi explained validity in the sense of accurate interpretation of the test score. Unlike most other large-scale standardized language tests, which are either norm-referenced or criterion-referenced, the CET-4 is defined as a criterion-related norm-referenced test in terms of score interpretation (http://www.cet.edu.cn/cet2011.htm). It is a norm-referenced test in the sense that the reported test score indicates the test-taker’s percentile position in the norm group. As reported in Section 4.1.1, the NCETC provides detailed explanations of the reported norm-referenced score in the test syllabus and on its official website. Scores of each administration were equated first using anchored students and then were normalized against a pre-determined norm group. Deyi highlighted the use of statistical analyses in the test development and equation processes to ensure validity of the reported norm-referenced CET-4 score. She said, “We have teams of program engineers, statisticians and psychometricians, who take care of equating the test, which is a very strict process. Results from (statistical item) analyses provide feedback regarding item difficulty and discrimination for test development.”

The criterion that the CET-4 is related to is the College English curriculum (i.e., the CECR). Deyi related this criterion to the purpose and nature of the test, saying that the CET-4 is an educational assessment in nature, which determines that the test should be based on the curriculum. The CECR provides requirements of English proficiency in terms of listening, speaking, reading, writing, and translation at three levels: the basic requirements, the intermediate requirements, and the higher requirements. These requirements specified learning outcomes for students at different levels. The CET-4 is based on the basic requirements (i.e., the minimum level that all non-English majors in
Chinese colleges and universities are expected to reach before graduation). Table 4.3 presents these requirements which I have translated from the Chinese version of the CECR. As the CET-4 under this study does not measure speaking, this is not included in the table.

Table 4.3

*Basic Level Requirements in the College English Curriculum*

| Listening | • Be able to follow classroom instructions, everyday conversations, and lectures on general topics conducted in English;  
• Be able to understand English radio and TV programs spoken at a speed of about 130 to 150 words per minute (wpm), grasping the main ideas and key points;  
• Be able to employ basic listening strategies to facilitate comprehension. |
| --- | --- |
| Reading | • Be able to read English texts on general topics at a speed of 70 wpm. With longer yet less difficult texts, the reading speed should be 100 wpm;  
• Be able to do skimming and scanning;  
• Be able to read textbooks in their areas of specialty with the help of dictionaries, and newspaper and magazine articles on familiar topics, grasping the main ideas and understanding major facts and relevant details;  
• Be able to understand texts of practical styles commonly used in work and daily life;  
• Be able to employ effective strategies in reading. |
| Writing | • Be able to complete writing tasks for general purposes, e.g., describing personal experiences, impressions, feelings, or some events, and to undertake practical writing;  
• Be able to write within 30 minutes a short composition of over 120 words on a general topic, or an outline. The composition should be basically complete in content, clear in main idea, appropriate in diction and coherent in discourse;  
• Be able to have a command of basic writing strategies. |
| Translation | • Be able to translate essays on familiar topics from English into Chinese and vice versa with the help of dictionaries;  
• The speed of translation from English into Chinese should be about 300 English words per hour whereas the speed of translation from Chinese into English should be around 250 Chinese characters per hour;  
• The translation should be basically accurate, free from serious mistakes in comprehension or expression. |
The requirements presented in Table 4.3 define the target domain of the CET-4. This target domain specifies the kinds of tasks and the ranges of contexts and conditions of observation in the CET-4. The test specifications are based on and consistent with these requirements. Deyi provided evidence for the connections between the curriculum, the test specifications, and the test design:

The CET-4 is based on the CECR. To make the CET-4 more accessible, we have published our test syllabus and sample test papers. In our test syllabus, we have test specifications based on the CECR. For example, for writing, the test-takers are required to write a passage of 120 words within 30 minutes. For listening, we have specified the types of materials used and the speed, etc. Also we have a vocabulary list. After so many years, teachers and students alike are very familiar with the test contents.

Considering the alignment between the test and the curriculum as well as the intended purposes of the test, I assumed that the CET-4 is an achievement test. However, Deyi stressed in the interview that the CET-4 was a norm-referenced language proficiency test rather than a criterion-referenced achievement test:

Strictly speaking, the [CET-4] test is not an achievement test because for an achievement test, there have to be very specific learning objectives and it is usually criterion-referenced. In general, the CET is [an English language] proficiency test. We specify our testing objectives with the guidance of the CECR. It [the CET-4] is a norm-referenced test in terms of score calculation and reporting. A proficiency test is usually norm-referenced.
The above explanation is consistent with the test construct defined in the documents disseminated by the CET-4 developer. For example, in the stated purposes of the test, the label *English language proficiency* is used to define the CET-4 construct. This is also consistent with the goals of College English education stated in the CECR (i.e., to develop students’ *comprehensive ability in using English*).

Considering the inseparability of language proficiency and language use context, I asked Deyi to explain the language use context in relation to the CET-4 construct. She said:

The CECR gives prominence to students’ English language use abilities. The construct measured by the CET-4 is students’ English language use abilities in listening, reading, and writing as specified in the CECR. The construct is students’ English language use abilities, not just knowledge. Language use ability needs to be interpreted in relation to context. The Reading Comprehension section of the CET-4 measures student’s reading ability in the academic context and the Writing and Listening Comprehension sections measure their English writing and listening abilities in general. So I think what the CET-4 measures is between EAP [English for academic purposes] and EGP [English for general purposes].

Deyi’s explanation suggests that the construct measured by the CET-4 spans across the academic context and the general English use context. The CECR also provides a broad description of the construct *comprehensive ability in using English*, which includes English communicative abilities in the students’ future studies and careers as well as social interactions. However, there is inconsistency between this broad description of the test construct and the narrow target domain based on which the test is
designed. This inconsistency constitutes construct underrepresentation threatening the
validity of the CET-4 score interpretation. Moreover, as the purpose of the norm-
referenced score is to compare students with each other, this kind of score reported in the
CET-4 would not provide much information about students’ achievement of the
curriculum. This would, in turn, threaten the utility of the test in terms of informing
College English teaching and learning.

4.2.2 Procedures Are Taken to Address Test Consequences
The CET-4 is intended not only to measure students’ English proficiency but also
to promote College English education in China. Therefore, consequences, particularly
consequences on English language teaching and learning are a fundamental consideration
of the CET-4 developer.

In the CET-4 context, the issue of consequences has been approached from two
perspectives: the test design perspective and the policy perspective. From the test design
perspective, overlap between the CET-4 and the curriculum is considered an important
factor to induce positive washback. In the interview, Deyi elaborated the consistency and
overlap between the test and the CECR as well as the implications of this overlapping for
washback:

The CECR and the CET reform are parts of the same quality project [College
English Reform Program, CERP]. Some of the people involved in developing the
curriculum are our NCETC members. So the CET is aligned with the curriculum
and the overlapping [between the CET and the curriculum] makes it more
possible for positive washback to be induced.
In order to induce positive washback and reduce negative washback of the CET-4, the NCETC made continuous efforts to revise the content and format of the test. As reported in Section 4.1.2, the proportion of MCQs was reduced over the years. Discrete-point items assessing knowledge about language such as grammar and vocabulary were replaced with a number of constructed-response item types.

Despite these efforts, negative consequences of the CET-4 remained a critical challenge to the test developer. In the interview, Deyi expressed concerns over the negative washback of the CET-4 and highlighted the dilemma facing the test developer. On the one hand, the NCETC intended to reduce the negative washback through improving the test design. On the other hand, the large scale of the test as well as its impact constituted limitations on test design improvement. Deyi said:

We are aware of the negative washback of the CET on students’ learning. For example, some students may prepare for the CET writing by memorizing some model essays or lengthy chunks of text. We cannot deduct their marks just because we suspect that they [students] reproduced the memorized chunks, if the overall writing is coherent and cohesive. What we can do to address this issue is improve the test design. You know, a test is like a knife, after some time it can become blunt. But we can only make limited changes in terms of test design; we cannot make dramatic changes. Otherwise, we would be overwhelmed by complaints and criticism. You know, for large-scale testing, there are a great many of constraints and we have to strive to survive and compromise with various competing forces.
At the policy level, recent reform of the CET-4 focused on reforming the test score reporting system to enhance the relevance of the test to College English education. The passing certificate was replaced with the score report showing both the total score and the sub scores for each section of the test. In order to enhance the utility of the CET-4 score for policy-making, the test developer explored ways in score reporting to provide useful information to educational policymakers at various levels. Item analysis data are reported to teachers to help them better understand their students’ strengths and weaknesses for future teaching. The writing score distribution data are given to institutions, indicating the average writing ability of the students in a given university or college. An Average Graded Score is used to indicate the overall proficiency of students in a university, a city, or a province.

The CET-4 developer also took some policy measures to address unintended consequences of the test in the educational and the societal contexts. For example, to mitigate the high stakes for students associated with the use of the CET-4 score for graduation and/or degree conferment, CMoE proposed an assessment model. The model emphasizes that the CET-4 score should be used together with evidence from other sources to make comprehensive evaluation of students’ English proficiency. Institutional authorities (i.e., individual universities or colleges) are encouraged to use the information provided by the CET-4 flexibly, together with information from other assessment sources, in making decisions in their particular contexts. To address the unintended consequences associated with the misuses and abuses of the CET-4 score in society, only students currently enrolled in higher educational institutions are eligible to take the CET-4. It is
also expected that by removing the passing certificate, the undue social pressure associated with the misuses/abuses of the certificate could be reduced.

The goal of promoting College English education by implementing the CET-4 testing program brought the issue of test consequences to the foreground. Results reported in this section revealed the continuous efforts of the CMoE and the NCETC to increase the positive consequences of the CET-4 and to decrease its negative impact. These efforts focus on improving the CET-4 design as well as formulating and implementing policies regarding College English assessment and CET-4 score reporting. However, the effectiveness of these policies is an issue that needs to be investigated by collecting evidence about how the test is actually being used by test users in both the educational and societal contexts. The following section reports results on the actual uses of the CET-4 results in these contexts.

4.3 Test Uses in Educational and Societal Contexts

Interview data with the eight test user participants in this study were analyzed inductively using thematic analysis (King & Horrocks, 2010) to understand the actual uses of the CET-4 in educational, business, and government contexts. These participants included two participants in the educational context (Edi and Edu); two participants in the business context where English was relevant in the workplace (Buran and Buru); two participants in the business context where English was not relevant in the workplace (Buna and Buni); and two participants in the government (Gongli and Gongfei) (see Table 3.1 for more information about these participants). The purpose of including these participants is to demonstrate the perspectives of test users in different contexts. Table
4.4 summarizes the categories concerning the actual uses of the CET-4 score with their codes and frequencies in relation to different contexts.

Table 4.4

*Matrix of Findings on Uses of CET-4 Results*

<table>
<thead>
<tr>
<th>Focus of research</th>
<th>Major findings</th>
<th>Frequency by context</th>
<th>Total frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ED*  BUR*  BUN*  GO*</td>
<td></td>
</tr>
<tr>
<td>Overarching Category: CET-4 results are used for different purposes</td>
<td>Codes:</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-is used as prerequisite for degree conferment to motivate students in education</td>
<td>7 - - -</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>-is used to compare students’ performance for promoting College English programs in education</td>
<td>5 - - -</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-is used as a threshold in employment recruitment for screening candidates</td>
<td>- 7 6 2</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>-is used by government to allocate social resources for future development of the city</td>
<td>- - 3 -</td>
<td>3</td>
</tr>
</tbody>
</table>

*Note. ED=educational context; BUR=business context where English was used in the workplace; BUN=business context where English was not used in the workplace; GO=government context*

4.3.1 Test Uses in the Educational Context

Both Edi’s and Edu's universities used the CET-4 total score of 425 as a cut-off score for students to obtain Bachelor’s degrees. In addition to the CET-4, internal university-based English tests were also used in these two universities.
In Edi’s university, students were required to take an English language placement test upon entering into their programs of studies. “New students came to our university with very varied levels of English,” said Edi in explaining the reason for the placement test, “and we need to place them to different classrooms so that students can be taught according to their different English levels.” The placement test placed the students into three levels: primary, medium, and high. Students at the primary level started from the Band 1 College English course, and they were not qualified to register for the CET-4 test until they completed the Band 4 College English course in the second year in the program. These students must take the CET-4 and have a CET-4 score over 425 to obtain their degrees. According to Edi, this group of students generally had less interest in studying English and that the CET-4 was useful in terms of keeping them working hard on English learning:

I have seen so many students in my classroom who worked very hard on English before [they took] the [CET-4] test. But once they have taken the test, they don’t listen to the teacher at all in the classroom. They think that the classroom-based test is not as important as the CET-4. You know this is a big issue in our university. Many teachers are complaining, and some suggested that students should not be allowed to take the test until after they complete the College English course.

Students who reached the medium level in the placement test at Edi’s university were qualified to take Band 3 College English course in the first semester and Band 4 in the second semester of their first year undergraduate studies. They could take CET-4 after taking these two courses in the first year in the program. They did not have to have a
CET-4 score to obtain their degrees. However, almost all these students would take the test and very few, if any, of them had a score below 425. Students at the high level of the placement test were qualified to be exempt from the College English course and the CET-4 score was not required to obtain their degrees. But again, almost all these students would take the test and have a score above 425. “The CET test is useful in the society anyway,” said Edi in explaining why these students wrote the CET-4 even though it was not required by the university.

In Edu’s university, there was no internal English placement test. All the undergraduate students took the College English courses Band 1 through Band 4 and took the CET-4 test after completing the Band 4 course in the second year in the program. Students needed to have a CET-4 score above 425 to obtain their academic degrees. In explaining this requirement, Edu pointed to the motivational function that large-scale high-stakes tests such as the CET-4 play in the Chinese educational context:

As everybody knows, Chinese education is still exam-driven. I don’t mean this is good, but this is the reality. No exam is perfect, but without the test, without using the test result to make decisions about students, I believe they would not be motivated as much to study English.

Students at Edu’s university who were not able to pass the cut-off score (CET-4 425) by the end of their 4-year program would have to pass an internal English exit test to obtain their academic degrees. According to Edu, the percentage of the students taking the exit test in his university was approximately 5%, and very few of them failed in this test. The intended purpose of using the internal exit test was to mitigate the high stakes of using the CET-4 for academic degree conferment. Edu said:
The major purpose of using the CET-4 as a prerequisite for degrees is to motivate students to learn. But we don’t want our students to graduate after four years of studies without a degree. So, we give the students (who did not pass the CET-4 cut-off score of 425) another chance to obtain their degrees by writing the internal College English test.

The uses of the CET-4 score for academic degree conferment in the two universities highlight the motivational role the test plays in the educational context. Used for high-stakes decisions, the test provides a practical and convenient tool for the policy makers to motivate the students to learn English. The motivational function is more related to the high stakes of the test than the test design. In other words, any high-stakes test may motivate learning in the Chinese educational context regardless of its design.

In Edi’s university, the motivational function of the CET-4 goes beyond motivating students’ learning. She highlighted the positive consequences of the CET-4 in increasing the financial input into the College English education program:

I think the CET plays a very important role in promoting the status of our program because students’ performance on the test is related to the reputation of the university. The CET is a national test, so universities can compare their students’ performance with each other. Students’ performance is also related to their employment after graduation, which in turn is related to the university’s reputation again. And students’ performance is an important consideration in quality evaluations of universities across the country. Therefore, in recent years, our university has put in large amount of money to College English education,
(e.g., to purchase more sophisticated teaching facilities such as language labs, etc.).

The above consequences highlight the indirect outcomes of the CET-4 program as policy implementation. These consequences are more related to the contextual factors operating on the Chinese educational scene.

In addition to the motivational function, Edi and Edu also pointed to the informational function of the CET-4. Their comments on this function focused on how the results of the test were used to know their students’ strengths and weaknesses in listening, reading, and writing. For example, Edu talked about the beneficial consequence of the CET-4 on their College English education program from the perspective of using the information provided by the test to enhance the quality of the program:

The CET makes it possible for us to compare our students’ performance with other universities and across the years. These comparisons enable us to have a better understanding of our strengths and weaknesses. For example, compared with students in other universities, our students are good at reading and writing, but listening has been our weakness in general. In the past few years, we have invested more money to build up new language labs. Listening has been given more importance in the classroom teaching. Our scores [on listening] are getting better now. So I think the CET is useful in providing feedback to College English teaching.

In Edi’s university, teachers in her department met at the beginning of each semester to discuss the reported CET-4 results and to make their instructional strategies
accordingly. Edi described how the CET score was used in her department to inform the College English education program:

In general, our students’ performances on the CET have been consistent over the years. After each CET administration, when we receive the test results from the NCETC, I will take a close look at them to see if there is any remarkable change in the score, not only the total score, but also the sub-scores of listening, reading, writing and so on, not only the university-level scores, but also scores related to each teacher. If I find some significant increase or decrease of the test scores, I will talk to the related teachers to try to find out the possible causes. If the increase or decrease is across the classrooms, we may call for a meeting.

In explaining the use of the CET-4 to inform teaching, Edi highlighted the relevance of the reported CET-4 score, particularly the sub-scores, to classroom teaching:

The reported CET scores, including the sub-scores, provide us with relevant information for classroom teaching. Compared with the old way of reporting the CET score, I like the current score reporting system much better because it provides more relevant information for teaching. For example, with the current system, it is possible for us to know how our students did in particular sections such as listening and writing, and we could make changes in our classrooms accordingly.

The informational function of the CET-4 was consistent with the CET-4 developer’s intended use of the test. However, this use in these two universities is based on comparison of students’ performances on the test with each other. The norm-
referred CET-4 score does not provide much information as to students’ achievement of the curriculum.

Edi and Edu also highlighted the negative washback effects of the CET-4, attributing them to the limitations of the test. They believed that the CET-4 underrepresented English language proficiency as defined in the College English curriculum, and expressed their concerns over “teaching and learning to the test”:

The new curriculum gives importance to student’s English use abilities, particularly in listening and speaking… But a test is a test. It has many constraints…So I think CET is a limited representation of students’ proficiency…

You know the writing section of the CET has a fixed format. This would encourage students to prepare for the test by rote learning some model essays and useful sentences. Memorization is perhaps important for English learning, but rote learning is not the best strategy to learn English. We need to learn by doing, by using English in the real context. I think the students need more practice in practical writing, for example, how to write an application letter and so on (Edu).

4.3.2 Test Uses in the Societal Contexts

My investigations of the uses of the CET-4 in the societal context focused on how the test score was used in the business and government contexts.

103
In the business context, the CET-4 score was used to recruit employees from universities. In Buna’s and Buni’s companies, English was not used in the workplace. However, a CET-4 score of over 425 was one of the qualifications on their recruitment advertisements. Both Buna and Buni considered this was the cut-off score for students to pass the CET-4 although the test developer had eliminated the pass-fail conception. Buna believed that this was a very low threshold for the job candidates, saying, “if you graduate from a prestigious university, it would be quite impossible for you to fail in the CET-4 test. A candidate with a low CET-4 score would have no opportunity for an interview unless we are convinced by other evidence that his other abilities are outstanding.”

Buna and Burni explained that the major reason for them to use the CET-4 score in the recruitment process was the large number of job applicants, which meant a heavy workload for the Human Resource Department. In order to reduce the workload, they used the CET-4 score as a gatekeeper or a threshold to screen out unqualified resumes. As Buna said,

Each year, we recruit approximately 30 university students, and receive about 500 applications. Each year too many students are graduating from the university, and we have limited positions for them. Think about the pressure of the employment market. We have a limited number of people working in the Human Resource Department; we are not able to deal with so many applications. Therefore, we need to set some bottom line for screening, and CET-4 425 is the minimum requirement for all undergraduate students.
Similarly, in Buni’s company, a lower CET-4 score would deny the applicant the opportunity of an interview while higher scores on international English language tests could give job applicants the edge in the competition.

In Buran’s and Buru’s companies, where English was used in the workplace, the CET-4 score was used for employee recruitment together with information about students’ English proficiency from other sources. Buran emphasized the flexibility in the use of the CET-4 score in the recruitment process:

In the process of recruitment, however, we just use the CET-4 score as a reference. That is to say, if you have your CET-4 score on your CV, and it’s over 425, that’s good. But we will not screen you out simply because you do not have the CET-4 score on your CV. You can provide other information to convince us of your English proficiency, for example, your experiences of using English, the English competitions you have participated in, your score on other accredited English tests such as TOEFL or IELTS, and so on.

Buran reiterated her interpretation of the CET-4 results in explaining why they collected evidence regarding the candidate’s English proficiency from other sources:

The CET score reflects the student’s learning in the classroom, the knowledge they learned from textbooks in the university. It’s academic English. What is used in the workplace is business English. So if a candidate has a BEC score, that would be evidence more relevant to the workplace. We are more interested in this type of English, that is, the student’s ability in using English in the business context.
In Buran’s company, shortlisted candidates were usually interviewed in English, in most cases with English-native speakers. “We can best assess the students’ English communicative ability in the interview in this way,” she said. In Buru’s company, there was an English test for shortlisted candidates. In this test, the candidates were required to complete different tasks based on the job positions they were applying for. These tasks included reading a job-related document, writing an email, a memo, or a short report based on the materials provided, and orally answering some related questions.

Buru highlighted the importance of collecting evidence from the internal test administered by the company:

In our recruitment, we use CET-4 425 as a qualification. But this does not mean that if your CET score is over 425, your English is sufficient for the job. Because the ability of using English is multidimensional. We need to consider your English in a comprehensive way. That is, we need to have evidence from different sources to have a more accurate evaluation of your English ability, and would weigh the evidence. Different evidence has different weighting. The CET-4 score is not as important as the result on our own test because our test assesses the candidates’ workplace English, and this is more relevant to the work. A student with a high CET-4 score may be good at reading textbooks in the university.

With regard to the CET-4 uses in the government context, in City A where Gongli worked, undergraduate students originally from other provinces or cities needed to apply for a residence permit called Hukou to be a permanent resident and work in the city under the local social insurance plan. The CET-4 score was required for this application.
Students’ merits were scored using a scoring criterion that includes factors such as CET score, GPA, extracurricular activities, as well as the prestige of the university where they graduated. When explaining why The CET-4 score was used in the scoring criterion for a Hukou application, Gongli highlighted the large population of the city and the large number of applicants:

The current population of this metropolis is over 23 million. 39% of the people living in the city are from outside. On the one hand, the increasing population has placed tremendous pressure on public services such as healthcare, transportation, and educational services. On the other hand, we want to attract more excellent talents to work in this city. Therefore we need a policy for regulation…. Every year there are a huge, an overwhelming, number of university graduates applying to work in this city. These students are not only from the city itself, but also from outside. The reason is that this city has more employment opportunities, higher income, and a better social insurance system than any other cities in the country. This is of course a good thing, but on the other hand, it presents a great challenge to us in terms of the limited social resources.

In contrast, City B, where Gongfei worked, was not as desirable to university graduates as City A in terms of employment. Students were not required to submit their CET scores to apply for a Hukou in this city. According to Gongfei, the majority of undergraduate students coming to live and work in the city were originally from it and had a Hukou registered there. “For those undergraduate students from other places, it’s not hard for them to obtain a Hukou here if they are hired by a local employer,” said

---

5 See Chapter 1 for more information about Hukou.
Gongfei, “we need to attract more university graduates to work in our city.” According to Gongfei, job positions in the city government were most desirable by most university graduates. The recruitment process for these positions involved three stages: 1) resume reviewing; 2) paper-and-pencil examination; and 3) face-to-face interview. In the first stage, applicants were required to submit their CET-4 scores together with their resumes. Applicants with a score below 425 would not have an opportunity for the next stage.

Thus, as in Buna’s and Buni’s companies, where the CET-4 score was used to screen out job candidates in the competitive recruitment process, the test score played a gatekeeping function in the government context in terms of the allocations of social resources and employment opportunities.

Overall, the results reported in this section highlight the multiple uses of the CET-4 in the business and government contexts. The test score is used to achieve different purposes by test users in the local contexts. These uses are high stakes for students. These findings reveal the power of the CET-4, which is beyond the control of the test developer.

4.4 Value Implications Underlying Uses of the CET-4 Results

The results reported above reveal the high stakes of the CET-4 for students. I believe that high-stakes decisions need to be justified based on the relevance between the test construct and the test use context. Therefore, a focus across the interviews with the eight test user participants is to find out how they justified their test use decisions. Analysis of the interview data revealed the value implications underlying test uses. Table 4.5 summarizes the categories concerning the values and their related codes.
Table 4.5

Matrix of Findings on Value Implications Underlying Test Uses

<table>
<thead>
<tr>
<th>Focus of Research</th>
<th>Major findings</th>
<th>Frequency by context</th>
<th>Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the value</td>
<td>Category 1: values associated with the construct measured by the CET-4</td>
<td>ED* BUR* BUN* GO*</td>
<td></td>
</tr>
<tr>
<td>implications underlying</td>
<td>Codes:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the actual uses of the CET-4?</td>
<td>- effort</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- quality of university graduates</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- basic requirement for all university graduates</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- talent with international visions and abilities</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Category 2: values associated with the test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Codes:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- credibility</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- reliability</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- authority</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ED=educational context; BUR=business context where English was used in the workplace; BUN=business context where English was not used in the workplace; GO=government context

4.4.1 Values Associated with the CET-4 Construct

In order to justify their test use decisions, test users in this study extended the test developer’s intended interpretation of the CET-4 score (i.e., representation of students’ English proficiency based on the College English curriculum). Analysis of the justifications these test users provided revealed the value implications underlying their extended interpretations.

The test users associated the CET-4 test construct with different traits and qualities of students. For example, in the government context, Gongfei and Gongli
associated the CET-4 score with talents with international visions and abilities. In explaining why the CET-4 score was considered important in recruiting employees for positions in the government, Gongfei expanded the interpretation of the CET-4 score from representation of English language proficiency to indication of cross-cultural communication competence and international visions. He said, “The CET score represents a university student’s English proficiency and English proficiency is part of cross-cultural communication competence. The development of our city needs young people with international visions and cross-cultural communication abilities. Language is part of the culture.”

Similarly, Gongli explained the use of the CET-4 score for Hukou applications in City A by relating the CET-4 score to “young people with a global vision” and highlighting the beneficial consequence in terms of the future development of the city:

Including the CET result in the criterion is based on the requirement for the development of the city. An international city needs young people with a global vision, and English is the language most often used in international communications. A university student with a global vision should have good English proficiency. We use the CET result in the scoring criterion in the hope that young people from outside the city with higher levels of English proficiency would have more opportunity to obtain a Hukou in here. This is beneficial to the future development of the city.

The extended interpretation of the CET-4 score in terms of talent with international visions and abilities goes beyond the target domain of the test. However, this interpretation is consistent with the description of College English education in
documents disseminated by the CMoE. For example, at the beginning of the CECR, College English is defined not only as a linguistic course, but also a course intended to expand students’ *international visions and international cultural knowledge* so as to meet the needs of China’s social development and international exchanges. Gongli related his extended interpretation to the revisions of the College English curriculum and the reform of the CET-4 test in recent years:

In order to accommodate the requirement for talents with high English proficiency and cross-cultural communication competence in society, the English education in Chinese universities has gone through significant reforms. These include revisions of the curriculum and reforms in the CET test. More importance is given to students’ English use abilities in international communications. The CET score is an indication of students’ competence in international communications.

To justify their test use decisions, Gongfei and Gongli not only associated the CET-4 score with traits such as international visions and abilities, but also related the score to student quality and the basic requirement for university graduates in China:

To pass the CET-4 is a basic requirement for all the university students…. A good command of English is part of the quality of modern university graduates. The CET-4 result is an indication of the quality of the graduate as it is required of all the university students…. To guarantee the personnel quality in the government, we need to recruit high quality university graduates. And the quality of the personnel working in the government is very important for the city (Gongfei).
There has been much controversy in recent years over the quality of higher education in our country and quality of the university graduates. The number of university students has increased dramatically. In this context, quality evaluation is particularly important. The CET is not only a test of the students’ English level, but also represents the quality; it’s a measurement for quality purpose. English is an important dimension of the quality of modern university graduates (Gongli).

The argument that the CET-4 is a proxy of student quality is consistent with results from my analysis of documents disseminated by the CMoE. In these documents, College English education is considered an important part of higher education in China. The implementations of the curriculum and the CET-4 programs served the goal of promoting the quality of College English education. For example, in the statement of the purpose of the CECR (CMoE, 2007a), College English education is related to quality of higher education and quality of students:

In order to accommodate the new development of higher education in China, deepen educational reform, improve educational quality, and to satisfy the needs of the country and society for qualified personnel in the new era, College English Curriculum Requirements have been drawn up to provide colleges and universities with guidelines for English instruction to non-English major students (my translation, with italics added for emphasis).

In documents disseminated by the CMoE, EFL is given policy priority. The CET-4 is a significant contributing factor to promoting the status of EFL in China. In this context, the power of the test is employed by test users to justify their test use decisions, and these decisions make the status of the test more powerful.
In the business context, English was relevant to the workplace in Buran’s and Buru’s companies where the CET-4 score was used for employee recruitment. However, the justifications they provided for their uses are not based on the relevance of the score to the English use context. Buran and Buru believed the test measured academic/school/textbook English rather than abilities to use English in the workplace context. Like Gongfei and Gongli in the government context, they sought to justify their uses of the CET-4 score by arguing that the test is a basic requirement for all university graduates. Buru said, “We cannot say a student with a high score has a very high level of English, but if his score is lower than 425, we can say he does not meet the basic requirement.”

This argument is consistent with the purposes of the CET-4 stated by the test developer. In the most recent test syllabus (Syllabus for College English Test, 2006, p. 1), the purposes of the CET are defined as “to examine the English language proficiency of undergraduate students in China and ensure that Chinese undergraduates reach the required English levels specified in the College English Curriculum Requirements.”

In Buna’s and Buni’s companies, English was not relevant to the workplace. However, the CET-4 score was still required in the recruitment process. To justify this use, Buna and Buni not only associated the CET-4 score with a basic requirement for students but also argued that the test was part of the student’s academic performance and a proxy of their attitudes and efforts:

A CET-4 score of over 425 is a basic requirement for the university graduates. Everybody knows the importance of English for university students in China nowadays. It is a key to success in the 21st century…. The reason we require the
CET-4 score is that it is part of their academic performance in the university. I believe a student should pass the CET-4 test within four years of studies in the university as a qualified university graduate (Buna). English is required for all the university students in China. The CET result does not only represent the student’s English level, but also is a requirement for all the university students. The CET score represents students’ English proficiency, and English proficiency is an important part of the overall quality of a university graduate. It is required of all the students. If you didn’t pass the CET-4, you are not a qualified university graduate… I would assume that you didn’t put enough effort in your studies. This is a matter of attitude. And attitude is also important in the workplace. So a student’s CET score could reflect many things that are important in the workplace, too. (Buni)

The association between the CET-4 score and effort was also made by Edi and Edu in the educational context. To justify the uses of the CET-4 score for academic degree conferment in their universities, they both argued that the score reflected students’ effort and highlighted the motivational function of this use. For example, Edu said:

The reason that we make the CET-4 score of 425 a prerequisite for degree conferment is basically to motivate students to study English. You know, the examination plays a particularly important role in motivating students. This is true in China, not only for primary and secondary students, but also for university students. And this has been the case for hundreds of years. We don’t want our students to leave the university without a degree. If they work hard on English,

---

6 He meant a CET-4 score below 425.
they are most likely to pass the CET-4. Even if their CET-4 score may be below 425, they can pass the internal test in the university and will be able to obtain their degree before graduation if they work hard.

Test users in this study sought to justify their test use decisions based on their extended interpretations of the test score. These interpretations are not based on the target domain of the test and go beyond the test developer’s intended interpretation. High-stakes uses based on these interpretations constitute not only validity issues but also ethical issues. The extended interpretations and uses of the CET-4 score are embedded in the Chinese educational and sociopolitical contexts, to which the test is a significant contributing factor.

4.4.2 Values Associated with the CET-4

Values associated with the CET-4 refer to the social values expressed by test users in this study to justify their test use decisions. Results from the analysis of the interview data show that the justifications they provided go beyond the construct interpretations reported in the last section. In order to justify their decisions, test users in this study highlighted values associated with the CET-4 per se, such as its perceived quality, credibility, reliability, and authority. They used the term reliability synonymously with credibility. Their arguments for the credibility and reliability of the CET-4 are generally based on their perceived qualities of the test. For example, in the educational context, Ed argued for the reliability and credibility of the CET-4 based on the consistency between the test score and students’ performance in the College English classroom:

The CET is a standardized test. I have been involved in supervising and scoring the CET papers and know that they [the NCETC] have followed very strict and
scientific procedures in test administration and scoring. It is a scientific test. I have also been teaching College English and helping students prepare for the CET. I find that students’ CET scores are consistent with their performance in the classroom. So I believe the CET is a credible and reliable instrument.

Edu stressed authority (i.e., the CET-4 is a national test administered by the CMoE) and the large scale of the test in explaining his perceptions about the reliability and high quality of the test:

You know CET is such a big test, millions of test-takers each year. And it is a national test, organized by the C MoE. For such a test, the quality has to be guaranteed. Although the test has limitations in representing students’ English proficiency, it should be a high quality and reliable test.

In the government context, Gongfei’s and Gongli’s perceptions of the CET-4 qualities are highly consistent with the test developer’s argument that the CET-4 is a scientific, fair, objective, and accurate measurement. Like Edu in the educational context, they both related the quality of the CET-4 to its authority and long history:

The CET is a national test administered by the Ministry of Education, so we believe in its authority. The test has been in place for over twenty years. It has been proved to be an objective and credible measurement of the university graduates’ English proficiency. It provides us with a fair way to select the candidates. Of course, it is not the only way (Gongfei).

We are aware that there is much controversy in society around the CET. But I believe we do need a national test like this in China. Otherwise, how can we gauge the quality of College English education? After all, it is a national initiative
and over the many years, it has changed a lot to accommodate the needs for university graduates in society. As a national test by the Ministry of Education, I believe it is a scientific, reliable and objective measurement of the student’s English proficiency (Gongli).

In the business context, the CET-4 score was used mainly for recruitment purposes. The value of the CET-4 for the test users in this context is that it provides a “useful and fair” means to compare candidates. Buni pointed out that recruitment was a selection process for which “an objective and fair instrument” was needed to compare the candidates and that the CET-4 provided such an instrument as it was a national test for all the students. Buna said in the interview:

The CET is a national test and that is the only test that all the university students are required to write across the country. So it provides a useful, objective, and fair instrument to compare the candidates from different parts of the country.

Like test user participants in other contexts, Buna and Buni argued for the credibility and reliability of the CET-4 based on their perceptions of the test quality. They related the quality of the CET-4 to its impact and long history. Buna said in the interview, “Think about the influences of the CET over so many years in our country. This is such an influential test in China. It has great impact on students’ studies and employment. Therefore, the quality of the test should be guaranteed.”

In short, in order to justify their test use decisions, test users in this study argued for the reliability and credibility of the CET-4 based on their perceived quality of the test. Their perceptions of the test quality were generally based on the authority, large scale, and long history of the test. Therefore, the justification was based on the circular
argument. These results suggest that any national large-scale tests in China, regardless of the constructs they are designed to measure, may be used for different purposes by test users. These findings further reveal the hidden power of the CET-4 in the Chinese educational and societal contexts.

4.5 Chapter Summary

The implementation of the CET-4 goes beyond the measurement level to the policy level. It is a policy decision made by the CMoE intended to serve the educational and sociopolitical needs in the Chinese society. The validity of the decision needs to be evaluated in terms of its consequences—intended and unintended, direct and indirect. Results reported in this chapter show that the CET-4 developer (i.e., the CMoE and the NCETC) is paying increasing attention to the impact of the test in both educational and societal contexts and addresses this issue from technical and policy perspectives. However, according to Deyi, the multiple uses of the CET-4 score and the consequences associated with these uses are beyond the control of the test developer. They are a function of the test and the context where the test is implemented. These findings highlight the power of the CET-4 in the Chinese society. This power goes beyond the construct measured by the CET-4 and the technical quality of the test, as it is embedded in the political, social, cultural, and educational contexts.
Chapter 5 Washback of the CET-4 on Learning and Learning Outcomes

This chapter presents results from analyses of the questionnaire data and the CET-4 test score data collected in this study. These results answer RQ3:

How does the CET-4 influence students’ learning and learning outcomes?

The above research question focuses on washback of the CET-4 on students’ learning and learning outcomes. Washback is considered a dimension of test impact in the educational system. Results on washback provide evidence regarding the intended and unintended consequences of the CET-4 in the Chinese tertiary educational context.

To answer the above question, a washback model was proposed in Chapter 3. Based on previous washback research, this model hypothesizes the relationships among students’ perceptions of the demands and uses of the CET-4, their test-taking expectations, their test preparation practices, and their test performances. These hypothesized relationships are illustrated in Figure 5.1.

![Figure 5-1. Proposed CET-4 Washback Model](attachment:image.png)
Structural equation modeling (SEM) was performed to evaluate this washback model with two types of quantitative data: questionnaire data and test score data. Before the SEM analysis, principal axis factoring (PAF) was conducted to extract the factor structure of the measured variables. Results from the PAF on the questionnaire data and the test score data were used as construct indicators for the five latent variables in the proposed washback model as illustrated by Figure 5.1.

This chapter consists of three sections. Section 5.1 presents descriptive statistics as well as the PAF results for the questionnaire data. Section 5.2 summarizes results from analyses of the CET-4 score data, including descriptive statistics of the participants’ scores for each of the sections of the test, the PAF results for the CET-4 test score data, and descriptive statistics of the subconstructs measured by the test. Section 5.3 focuses on the SEM results, including the model fit indices and the parameter estimates, to evaluate the adequacy and appropriateness of the proposed CET-4 washback model.

5.1 Perceptions of CET-4 Demands and Uses, Test-taking Expectations, and Test Preparation Practices

5.1.1 Perceptions of CET-4 Demands

The first section of the questionnaire included 44 items measuring students’ perceptions of the demands related to different sections of the CET-4 (listening comprehension; reading comprehension; integrated skills; and writing and translation). Table 5.1 summarizes the descriptive statistics of these questionnaire items.
Table 5.1

Descriptive Statistics of Perceptions of CET-4 Demands

<table>
<thead>
<tr>
<th>Test sections</th>
<th>Questionnaire item</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening comprehension</td>
<td>LC1</td>
<td>5.69</td>
<td>1.71</td>
<td>-0.68</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>LC2</td>
<td>5.96</td>
<td>1.40</td>
<td>-0.70</td>
<td>-0.53</td>
</tr>
<tr>
<td></td>
<td>LC3</td>
<td>5.78</td>
<td>1.52</td>
<td>-0.61</td>
<td>-0.64</td>
</tr>
<tr>
<td></td>
<td>LC4</td>
<td>5.84</td>
<td>1.53</td>
<td>-0.68</td>
<td>-0.61</td>
</tr>
<tr>
<td></td>
<td>LC5</td>
<td>4.89</td>
<td>1.69</td>
<td>-0.24</td>
<td>-0.64</td>
</tr>
<tr>
<td></td>
<td>LC6</td>
<td>5.54</td>
<td>1.46</td>
<td>-0.38</td>
<td>-0.64</td>
</tr>
<tr>
<td></td>
<td>LC7</td>
<td>4.92</td>
<td>1.66</td>
<td>-0.20</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>LC8</td>
<td>5.53</td>
<td>1.49</td>
<td>-0.40</td>
<td>-0.65</td>
</tr>
<tr>
<td></td>
<td>LC9</td>
<td>5.50</td>
<td>1.49</td>
<td>-0.40</td>
<td>-0.64</td>
</tr>
<tr>
<td></td>
<td>LC10</td>
<td>5.24</td>
<td>1.57</td>
<td>-0.27</td>
<td>-0.64</td>
</tr>
<tr>
<td></td>
<td>LC11</td>
<td>4.59</td>
<td>1.57</td>
<td>-0.09</td>
<td>-0.47</td>
</tr>
<tr>
<td></td>
<td>LC12</td>
<td>5.96</td>
<td>1.34</td>
<td>-0.61</td>
<td>-0.56</td>
</tr>
<tr>
<td></td>
<td>LC13</td>
<td>5.68</td>
<td>1.42</td>
<td>-0.45</td>
<td>-0.66</td>
</tr>
<tr>
<td></td>
<td>LC14</td>
<td>5.43</td>
<td>1.39</td>
<td>-0.30</td>
<td>-0.58</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>RC1</td>
<td>6.05</td>
<td>1.40</td>
<td>-0.83</td>
<td>-0.44</td>
</tr>
<tr>
<td></td>
<td>RC2</td>
<td>6.15</td>
<td>1.21</td>
<td>-0.78</td>
<td>-0.42</td>
</tr>
<tr>
<td></td>
<td>RC3</td>
<td>4.96</td>
<td>1.61</td>
<td>-0.20</td>
<td>-0.60</td>
</tr>
<tr>
<td></td>
<td>RC4</td>
<td>5.93</td>
<td>1.31</td>
<td>-0.64</td>
<td>-0.56</td>
</tr>
<tr>
<td></td>
<td>RC5</td>
<td>5.97</td>
<td>1.36</td>
<td>-0.67</td>
<td>-0.58</td>
</tr>
<tr>
<td></td>
<td>RC6</td>
<td>5.97</td>
<td>1.35</td>
<td>-0.66</td>
<td>-0.56</td>
</tr>
<tr>
<td></td>
<td>RC7</td>
<td>6.14</td>
<td>1.22</td>
<td>-0.78</td>
<td>-0.44</td>
</tr>
<tr>
<td></td>
<td>RC8</td>
<td>5.95</td>
<td>1.31</td>
<td>-0.55</td>
<td>-0.57</td>
</tr>
<tr>
<td></td>
<td>RC9</td>
<td>5.19</td>
<td>1.52</td>
<td>-0.24</td>
<td>-0.57</td>
</tr>
<tr>
<td></td>
<td>RC10</td>
<td>5.83</td>
<td>1.48</td>
<td>-0.63</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>RC11</td>
<td>6.14</td>
<td>1.33</td>
<td>-0.85</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td>RC12</td>
<td>5.87</td>
<td>1.40</td>
<td>-0.58</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>RC13</td>
<td>5.73</td>
<td>1.35</td>
<td>-0.47</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>RC14</td>
<td>5.63</td>
<td>1.42</td>
<td>-0.41</td>
<td>-0.64</td>
</tr>
<tr>
<td>Integrated skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IS1</td>
<td>6.09</td>
<td>1.34</td>
<td>-0.75</td>
<td>-0.51</td>
</tr>
<tr>
<td></td>
<td>IS2</td>
<td>5.84</td>
<td>1.41</td>
<td>-0.57</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>IS3</td>
<td>5.17</td>
<td>1.48</td>
<td>-0.23</td>
<td>-0.53</td>
</tr>
<tr>
<td></td>
<td>IS4</td>
<td>5.85</td>
<td>1.37</td>
<td>-0.58</td>
<td>-0.61</td>
</tr>
<tr>
<td></td>
<td>IS5</td>
<td>6.15</td>
<td>1.25</td>
<td>-0.80</td>
<td>-0.43</td>
</tr>
<tr>
<td></td>
<td>IS6</td>
<td>5.68</td>
<td>1.48</td>
<td>-0.49</td>
<td>-0.65</td>
</tr>
<tr>
<td></td>
<td>IS7</td>
<td>5.92</td>
<td>1.34</td>
<td>-0.61</td>
<td>-0.58</td>
</tr>
<tr>
<td>Writing and translation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WT1</td>
<td>5.99</td>
<td>1.45</td>
<td>-0.77</td>
<td>-0.50</td>
</tr>
<tr>
<td></td>
<td>WT2</td>
<td>6.07</td>
<td>1.32</td>
<td>-0.76</td>
<td>-0.48</td>
</tr>
<tr>
<td></td>
<td>WT3</td>
<td>6.02</td>
<td>1.35</td>
<td>-0.69</td>
<td>-0.54</td>
</tr>
<tr>
<td></td>
<td>WT4</td>
<td>5.98</td>
<td>1.58</td>
<td>-0.90</td>
<td>-0.42</td>
</tr>
<tr>
<td></td>
<td>WT5</td>
<td>6.11</td>
<td>1.34</td>
<td>-0.82</td>
<td>-0.44</td>
</tr>
<tr>
<td></td>
<td>WT6</td>
<td>6.05</td>
<td>1.35</td>
<td>-0.70</td>
<td>-0.55</td>
</tr>
<tr>
<td></td>
<td>WT7</td>
<td>6.13</td>
<td>1.34</td>
<td>-0.82</td>
<td>-0.46</td>
</tr>
<tr>
<td></td>
<td>WT8</td>
<td>6.31</td>
<td>1.31</td>
<td>-1.19</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>WT9</td>
<td>6.40</td>
<td>1.22</td>
<td>-1.30</td>
<td>0.41</td>
</tr>
</tbody>
</table>
On a 7-point visual analog scale (1=strongly disagree; 7=strongly agree), the means of the item scores ranged from a low of 4.59 (LC11: background knowledge is important to answer listening comprehension questions correctly) to a high of 6.40 (WT9: my handwriting must be neat and tidy in order to do the writing section well). Compared with items measuring perceptions of other sections of the test, the means of the nine items on perceptions about the demands of the writing and translation section (WT1—WT9) were relatively higher, ranging from 5.98 (WT4) to 6.40 (WT9); the means of the items for the listening comprehension section (LC1—LC14) were relatively lower, ranging from 4.59 (LC11) to 5.96 (LC2 and LC12). There were four items with the means below 5.0 (RC3, LC7, LC5, and LC11). The standard deviations of the items were between 1.21 (item RC2) and 1.71 (item LC1). In general, items with higher means had smaller standard deviation values, indicating a ceiling effect. The skewness values of all the items were negative, indicating that more of the responses are on the higher end of the scale.

Two items concerning perceptions of the demands of the writing and translation section of the test had skewness values below -1.00 (WT8, -1.19; and WT9, -1.30). The kurtosis values of these two items are both positive (WT8, 0.19; WT9, 0.41). These statistics, together with the higher means of all the nine items on perceptions of the demands of the writing and translation section, indicated that students’ perceptions of the demands of this section are high. The kurtosis values of all the other items, except those of items WT8 and WT9, were all negative.

The KMO measure of sampling adequacy for the items in this section was 0.97. Bartlett’s test of sphericity was significant (p<.001). These statistics indicated that the variables were sufficiently correlated for a PAF to be appropriate. According to Comrey
and Lee (1992), loadings of .71 or higher can be considered “excellent”, .63 is “very good”, .55 is “good”, .45 is “fair”, and .32 is “poor.” Based on these criteria, items with poor loadings across the factors are referred to as outlying variables and items with fairly high loadings on two or more factors are called complex variables. Four outlying and complex variables (RC1, RC4, RC5 and IS6) were found and removed. After the removal, the remaining variables were re-analyzed. Varimax was used as the method for rotation. Rotated factor matrix from this analysis was examined to see if the factor structure had achieved structural simplicity and substantive meaningfulness. Table 5.2 presents the final PAF results.

To identify factors, two standards were set: 1) they obtained eigen values of 1 or above, and 2) they were located at or above a point of inflexion (or ‘elbow’) on a scree plot. The PAF resulted in four factors concerning perceptions of CET-4 task demands. Eighteen items loaded highly on the first factor. These items were primarily about perceptions of the English language knowledge and skills required for answering the CET-4 questions. Thus this factor was labeled *English language knowledge and skills* (LKS). The 12 items that loaded highly on the second factor primarily concerned perceptions of demands to answer the listening comprehension questions on the test. Thus this factor was defined as *listening skills* (LIS). The five items that loaded highly on Factor 3 related to perceptions of background knowledge that is required for answering the test questions. This factor was referred to as *background knowledge* (BAK). Five items loaded highly on the fourth factor. These items were all about perceptions of demands to answer the reading comprehension questions. This factor was defined as *reading skills* (RES).
Table 5.2

*Rotated Factor Matrix of Perceptions of CET-4 Demands*

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT2</td>
<td>.766</td>
<td>.224</td>
<td>.155</td>
<td>.027</td>
</tr>
<tr>
<td>WT1</td>
<td>.743</td>
<td>.198</td>
<td>.162</td>
<td>-.066</td>
</tr>
<tr>
<td>WT5</td>
<td>.739</td>
<td>.261</td>
<td>.202</td>
<td>.180</td>
</tr>
<tr>
<td>WT6</td>
<td>.710</td>
<td>.291</td>
<td>.210</td>
<td>.112</td>
</tr>
<tr>
<td>IS1</td>
<td>.710</td>
<td>.269</td>
<td>.159</td>
<td>.140</td>
</tr>
<tr>
<td>WT8</td>
<td>.710</td>
<td>.199</td>
<td>.068</td>
<td>.285</td>
</tr>
<tr>
<td>WT9</td>
<td>.683</td>
<td>.270</td>
<td>.120</td>
<td>.209</td>
</tr>
<tr>
<td>WT4</td>
<td>.675</td>
<td>.263</td>
<td>.077</td>
<td>.207</td>
</tr>
<tr>
<td>WT7</td>
<td>.651</td>
<td>.288</td>
<td>.206</td>
<td>.184</td>
</tr>
<tr>
<td>WT3</td>
<td>.639</td>
<td>.258</td>
<td>.225</td>
<td>.104</td>
</tr>
<tr>
<td>RC11</td>
<td>.633</td>
<td>.334</td>
<td>.119</td>
<td>.282</td>
</tr>
<tr>
<td>IS2</td>
<td>.620</td>
<td>.287</td>
<td>.227</td>
<td>.119</td>
</tr>
<tr>
<td>IS5</td>
<td>.619</td>
<td>.292</td>
<td>.180</td>
<td>.114</td>
</tr>
<tr>
<td>RC7</td>
<td>.558</td>
<td>.338</td>
<td>.185</td>
<td>.365</td>
</tr>
<tr>
<td>IS4</td>
<td>.545</td>
<td>.323</td>
<td>.252</td>
<td>.105</td>
</tr>
<tr>
<td>RC2</td>
<td>.542</td>
<td>.312</td>
<td>.183</td>
<td>.260</td>
</tr>
<tr>
<td>RC12</td>
<td>.494</td>
<td>.324</td>
<td>.230</td>
<td>.226</td>
</tr>
<tr>
<td>IS7</td>
<td>.493</td>
<td>.331</td>
<td>.221</td>
<td>.117</td>
</tr>
<tr>
<td>LC8</td>
<td>.298</td>
<td>.695</td>
<td>.000</td>
<td>.194</td>
</tr>
<tr>
<td>LC2</td>
<td>.201</td>
<td>.670</td>
<td>.223</td>
<td>.323</td>
</tr>
<tr>
<td>LC6</td>
<td>.252</td>
<td>.615</td>
<td>.178</td>
<td>.194</td>
</tr>
<tr>
<td>LC4</td>
<td>.332</td>
<td>.570</td>
<td>.174</td>
<td>.049</td>
</tr>
<tr>
<td>LC12</td>
<td>.433</td>
<td>.564</td>
<td>.120</td>
<td>.167</td>
</tr>
<tr>
<td>LC3</td>
<td>.331</td>
<td>.555</td>
<td>.084</td>
<td>.079</td>
</tr>
<tr>
<td>LC7</td>
<td>.113</td>
<td>.547</td>
<td>.275</td>
<td>.133</td>
</tr>
<tr>
<td>LC13</td>
<td>.272</td>
<td>.546</td>
<td>.254</td>
<td>.132</td>
</tr>
<tr>
<td>LC9</td>
<td>.294</td>
<td>.546</td>
<td>.267</td>
<td>.138</td>
</tr>
<tr>
<td>LC5</td>
<td>.142</td>
<td>.528</td>
<td>.286</td>
<td>.053</td>
</tr>
<tr>
<td>LC14</td>
<td>.317</td>
<td>.522</td>
<td>.287</td>
<td>.185</td>
</tr>
<tr>
<td>LC1</td>
<td>.346</td>
<td>.495</td>
<td>.101</td>
<td>.099</td>
</tr>
<tr>
<td>RC3</td>
<td>.097</td>
<td>.190</td>
<td>.816</td>
<td>.129</td>
</tr>
<tr>
<td>IS3</td>
<td>.318</td>
<td>.134</td>
<td>.713</td>
<td>.055</td>
</tr>
<tr>
<td>LC11</td>
<td>.148</td>
<td>.338</td>
<td>.681</td>
<td>.077</td>
</tr>
<tr>
<td>LC10</td>
<td>.227</td>
<td>.323</td>
<td>.459</td>
<td>-.040</td>
</tr>
<tr>
<td>RC14</td>
<td>.332</td>
<td>.358</td>
<td>.298</td>
<td>.752</td>
</tr>
<tr>
<td>RC13</td>
<td>.415</td>
<td>.343</td>
<td>.278</td>
<td>.746</td>
</tr>
<tr>
<td>RC10</td>
<td>.383</td>
<td>.321</td>
<td>.269</td>
<td>.609</td>
</tr>
<tr>
<td>RC6</td>
<td>.378</td>
<td>.321</td>
<td>.142</td>
<td>.603</td>
</tr>
<tr>
<td>RC8</td>
<td>.405</td>
<td>.401</td>
<td>.220</td>
<td>.587</td>
</tr>
</tbody>
</table>

Summative factor scores were calculated by averaging the summing scores for all items representing a factor (DiStefano et al., 2009). Table 5.3 summarizes the descriptive statistics of the four factors concerning students’ perceptions of the CET-4 task demands,
including factor names, items loading highly on each of the factors, means and standard deviations of the factor scores.

Table 5.3

**Descriptive Statistics of Factors Concerning Perceptions of CET-4 Task Demands**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>English language knowledge and skills (LKS)</td>
<td>RC2, RC7, RC11, RC12, IS1, IS2, IS4, IS5, IS7, WT1, WT2, WT3, WT4, WT5, WT6, WT7, WT8, WT9</td>
<td>6.07</td>
<td>1.03</td>
<td>-1.13</td>
<td>1.09</td>
</tr>
<tr>
<td>Listening skills (LIS)</td>
<td>LC1, LC2, LC3, LC4, LC5, LC6, LC7, LC8, LC9, LC12, LC13, LC14</td>
<td>5.56</td>
<td>1.07</td>
<td>-0.70</td>
<td>0.55</td>
</tr>
<tr>
<td>Background knowledge (BAK)</td>
<td>LC10, LC11, RC3, RC9, IS3</td>
<td>5.03</td>
<td>1.23</td>
<td>-0.34</td>
<td>-0.08</td>
</tr>
<tr>
<td>Reading skills (RES)</td>
<td>RC6, RC8, RC10, RC13, RC14</td>
<td>5.82</td>
<td>1.13</td>
<td>-0.63</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Of the four factors, LKS had the highest mean (6.07) and the highest kurtosis value (1.09) while its skewness value (-1.13) and standard deviation (1.03) were the lowest. These statistics indicated students perceived that the demands of the CET-4 tasks in terms of English language knowledge and skills were high. The means of LIS (5.56) and RES (5.82) were also relatively high. BAK has the lowest mean (5.03) and the highest standard deviation, indicating that there was bigger variation in perceptions of the CET-4 task demand in terms of background knowledge.

### 5.1.2 Perceptions of CET-4 Score Uses

The second section of the questionnaire contained 16 items that measured students’ perceptions of the uses of the CET-4 score. These items were intended to have
three sub-scales: instrumental uses (7 items), achievement uses (5 items), and task values (4 items). Table 5.4 presents the descriptive statistics of these items.

Table 5.4

Descriptive Statistics of Perceptions of CET-4 Uses

<table>
<thead>
<tr>
<th>Item</th>
<th>$M$</th>
<th>$SD$</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU1</td>
<td>5.04</td>
<td>1.84</td>
<td>-0.30</td>
<td>-0.85</td>
</tr>
<tr>
<td>PU2</td>
<td>5.10</td>
<td>1.80</td>
<td>-0.28</td>
<td>-0.82</td>
</tr>
<tr>
<td>PU3</td>
<td>5.52</td>
<td>1.64</td>
<td>-0.43</td>
<td>-0.76</td>
</tr>
<tr>
<td>PU4</td>
<td>5.78</td>
<td>1.54</td>
<td>-0.58</td>
<td>-0.70</td>
</tr>
<tr>
<td>PU5</td>
<td>4.97</td>
<td>1.77</td>
<td>-0.20</td>
<td>-0.76</td>
</tr>
<tr>
<td>PU6</td>
<td>5.44</td>
<td>1.73</td>
<td>-0.44</td>
<td>-0.78</td>
</tr>
<tr>
<td>PU7</td>
<td>5.13</td>
<td>1.75</td>
<td>-0.29</td>
<td>-0.78</td>
</tr>
<tr>
<td>PU8</td>
<td>5.26</td>
<td>1.79</td>
<td>-0.39</td>
<td>-0.84</td>
</tr>
<tr>
<td>PU9</td>
<td>4.49</td>
<td>1.62</td>
<td>-0.06</td>
<td>-0.48</td>
</tr>
<tr>
<td>PU10</td>
<td>4.65</td>
<td>1.64</td>
<td>-0.11</td>
<td>-0.56</td>
</tr>
<tr>
<td>PU11</td>
<td>4.83</td>
<td>1.59</td>
<td>-0.13</td>
<td>-0.50</td>
</tr>
<tr>
<td>PU12</td>
<td>4.84</td>
<td>1.57</td>
<td>-0.13</td>
<td>-0.53</td>
</tr>
<tr>
<td>PU13</td>
<td>5.47</td>
<td>1.62</td>
<td>-0.40</td>
<td>-0.71</td>
</tr>
<tr>
<td>PU14</td>
<td>5.38</td>
<td>1.65</td>
<td>-0.37</td>
<td>-0.76</td>
</tr>
<tr>
<td>PU15</td>
<td>5.61</td>
<td>1.53</td>
<td>-0.46</td>
<td>-0.71</td>
</tr>
<tr>
<td>PU16</td>
<td>5.49</td>
<td>1.66</td>
<td>-0.45</td>
<td>-0.76</td>
</tr>
</tbody>
</table>

On a 7-point visual analog scale, the means of the item scores ranged from the lowest 4.49 (PU9: CET scores provide an accurate measure of my English proficiency) to the highest 5.78 (PU4: I take the CET-4 mainly to graduate with a degree). Compared with other items in this section of the questionnaire, the standard deviations of PU9 and PU4 were relatively small (1.62 and 1.54 respectively). These statistics indicated that perceptions of the CET-4 uses in terms of providing accurate information about English proficiency were relatively low while perceptions of the CET-4 uses in terms of degree requirements were high. The skewness and kurtosis values of all the items were negative, with none less than -1.00, indicating a relatively normal distribution of the data.

The KMO measure of sampling adequacy for items in this section was .93. Bartlett’s test of sphericity was significant ($p<.001$). These statistics indicated that it was
appropriate to use the PAF. The PAF resulted in three factors concerning students’ perceptions of the CET-4 score uses. The factor loadings of different items are summarized in Table 5.5.

Table 5.5

Rotated Factor Matrix of Perceptions of CET-4 Uses

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU12</td>
<td>.869</td>
<td>.189</td>
<td>.260</td>
</tr>
<tr>
<td>PU11</td>
<td>.847</td>
<td>.199</td>
<td>.281</td>
</tr>
<tr>
<td>PU10</td>
<td>.842</td>
<td>.290</td>
<td>.234</td>
</tr>
<tr>
<td>PU9</td>
<td>.792</td>
<td>.290</td>
<td>.276</td>
</tr>
<tr>
<td>PU5</td>
<td>.171</td>
<td>.708</td>
<td>.230</td>
</tr>
<tr>
<td>PU4</td>
<td>.092</td>
<td>.677</td>
<td>.196</td>
</tr>
<tr>
<td>PU2</td>
<td>.263</td>
<td>.648</td>
<td>.069</td>
</tr>
<tr>
<td>PU14</td>
<td>.304</td>
<td>.600</td>
<td>.331</td>
</tr>
<tr>
<td>PU7</td>
<td>.261</td>
<td>.582</td>
<td>.310</td>
</tr>
<tr>
<td>PU6</td>
<td>.062</td>
<td>.516</td>
<td>.331</td>
</tr>
<tr>
<td>PU1</td>
<td>.222</td>
<td>.467</td>
<td>.199</td>
</tr>
<tr>
<td>PU15</td>
<td>.266</td>
<td>.252</td>
<td>.801</td>
</tr>
<tr>
<td>PU16</td>
<td>.246</td>
<td>.226</td>
<td>.766</td>
</tr>
<tr>
<td>PU13</td>
<td>.294</td>
<td>.406</td>
<td>.634</td>
</tr>
<tr>
<td>PU8</td>
<td>.365</td>
<td>.312</td>
<td>.546</td>
</tr>
<tr>
<td>PU3</td>
<td>.275</td>
<td>.423</td>
<td>.510</td>
</tr>
</tbody>
</table>

The three factors were consistent with the intended sub-scales of the second section in the questionnaire. All the items loaded highly on the sub-scales they were designed to measure. Specifically, four items loaded on the first factor. This factor was referred to as task values (TAV) (i.e., students’ perceptions of the values of the CET-4 tasks in terms of measuring their English proficiency). The seven items that loaded highly on the second factor related to the instrumental uses of the CET-4 score. It was referred to as instrumental uses (INU). The third factor concerned students’ perceptions of the
CET-4 usefulness in terms of their English learning achievement. This factor was referred to as *achievement uses* (ACU). Five items loaded on this factor.

The items that loaded highly on each of these three factors and the descriptive statistics of the summative factor scores, including the means, standard deviations, and skewness and kurtosis values are presented in Table 5.6.

**Table 5.6**

*Descriptive Statistics of Factors Concerning Perceptions of CET-4 Uses*

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task values (TAV)</td>
<td>PU9, PU10, PU11, PU12</td>
<td>4.70</td>
<td>1.50</td>
<td>-0.16</td>
<td>-0.32</td>
</tr>
<tr>
<td>Instrumental uses (INU)</td>
<td>PU1, PU2, PU4, PU5, PU6, PU7, PU14</td>
<td>5.26</td>
<td>1.26</td>
<td>-0.32</td>
<td>-0.43</td>
</tr>
<tr>
<td>Achievement uses (ACU)</td>
<td>PU3, PU8, PU13, PU15, PU16</td>
<td>5.47</td>
<td>1.37</td>
<td>-0.45</td>
<td>-0.44</td>
</tr>
</tbody>
</table>

As shown by the statistics in Table 5.6, the mean of TAV (4.70) was low compared with INU (M=5.26) and ACU (M=5.47). The values of the skewness and kurtosis of all the factors were negative, with none lower than -.45, suggesting a relatively normal distribution of the data.

**5.1.3 Test-taking Expectations**

The third section of the questionnaire contained 21 items that measured students’ CET-4 test-taking expectations. These items were intended for two sub scales: students’ self-efficacy in their English proficiency in relation to the CET-4 demands (items EX1—EX11) and their perceptions of the CET-4 task difficulty (EX12—EX21). Table 5.7 presents the descriptive statistics of these items.
Table 5.7

*Descriptive Statistics of Students’ Expectations of Taking CET-4*

<table>
<thead>
<tr>
<th>Item</th>
<th>$M$</th>
<th>$SD$</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>EX1</td>
<td>5.61</td>
<td>1.57</td>
<td>-0.46</td>
<td>-0.76</td>
</tr>
<tr>
<td>EX 2</td>
<td>5.52</td>
<td>1.51</td>
<td>-0.37</td>
<td>-0.68</td>
</tr>
<tr>
<td>EX 3</td>
<td>5.12</td>
<td>1.56</td>
<td>-0.20</td>
<td>-0.59</td>
</tr>
<tr>
<td>EX 4</td>
<td>5.07</td>
<td>1.64</td>
<td>-0.22</td>
<td>-0.68</td>
</tr>
<tr>
<td>EX 5</td>
<td>5.09</td>
<td>1.45</td>
<td>-0.21</td>
<td>-0.57</td>
</tr>
<tr>
<td>EX 6</td>
<td>5.44</td>
<td>1.47</td>
<td>-0.30</td>
<td>-0.63</td>
</tr>
<tr>
<td>EX 7</td>
<td>5.41</td>
<td>1.42</td>
<td>-0.27</td>
<td>-0.58</td>
</tr>
<tr>
<td>EX 8</td>
<td>5.15</td>
<td>1.52</td>
<td>-0.24</td>
<td>-0.59</td>
</tr>
<tr>
<td>EX 9</td>
<td>5.28</td>
<td>1.47</td>
<td>-0.26</td>
<td>-0.58</td>
</tr>
<tr>
<td>EX 10</td>
<td>5.15</td>
<td>1.51</td>
<td>-0.21</td>
<td>-0.58</td>
</tr>
<tr>
<td>EX 11</td>
<td>5.27</td>
<td>1.45</td>
<td>-0.22</td>
<td>-0.59</td>
</tr>
<tr>
<td>EX 12</td>
<td>3.92</td>
<td>2.01</td>
<td>0.02</td>
<td>-0.83</td>
</tr>
<tr>
<td>EX 13</td>
<td>4.95</td>
<td>1.84</td>
<td>-0.27</td>
<td>-0.79</td>
</tr>
<tr>
<td>EX 14</td>
<td>5.11</td>
<td>1.77</td>
<td>-0.31</td>
<td>-0.76</td>
</tr>
<tr>
<td>EX 15</td>
<td>4.96</td>
<td>1.79</td>
<td>-0.21</td>
<td>-0.58</td>
</tr>
<tr>
<td>EX 16</td>
<td>4.30</td>
<td>1.81</td>
<td>-0.05</td>
<td>-0.60</td>
</tr>
<tr>
<td>EX 17</td>
<td>4.48</td>
<td>1.71</td>
<td>-0.08</td>
<td>-0.53</td>
</tr>
<tr>
<td>EX 18</td>
<td>4.84</td>
<td>1.63</td>
<td>-0.14</td>
<td>-0.52</td>
</tr>
<tr>
<td>EX 19</td>
<td>4.50</td>
<td>1.72</td>
<td>-0.07</td>
<td>-0.57</td>
</tr>
<tr>
<td>EX 20</td>
<td>4.88</td>
<td>1.66</td>
<td>-0.17</td>
<td>-0.58</td>
</tr>
<tr>
<td>EX 21</td>
<td>4.70</td>
<td>1.66</td>
<td>-0.12</td>
<td>-0.55</td>
</tr>
</tbody>
</table>

The means of the variables regarding students’ self-efficacy were all above 5.00, the lowest being the mean (5.07) of EX4 (self-efficacy in doing well in passage listening in the listening comprehension section) and the highest being the mean (5.61) of EX1 (self-efficacy in doing well in the CET-4). With regard to perceptions of the CET-4 difficulty, the means of the item scores ranged from 3.92 (EX12, perception of the difficulty of the short conversations in the listening comprehension section) to 5.11 (EX14, perception of passage listening in the listening comprehension section). In general, the standard deviations of items EX12—EX21 were higher than those of items EX1—EX11, indicating that there is greater variation in students’ perceptions of the CET-4 difficulty and less variation in their self-efficacy in doing well in the test.

Statistics in Table 5.7 also suggest an association between students’ self-efficacy and
their perceptions of test difficulty. In general, tasks in which students had higher self-efficacy were perceived as less difficult and vice versa.

The skewness and kurtosis values of all the items, except the skewness of EX12, are negative, but none was smaller than -1.00, indicating a relatively normal distribution of the data. The KMO measure of sampling adequacy for the items in this section was .89. Bartlett’s test of sphericity was significant ($p<.001$), indicating that it was appropriate to use the PAF. The PAF resulted in two factors, which were consistent with the intended sub-scales of these variables. The factor loadings of different items are presented in Table 5.8.

Table 5.8

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EX 3</td>
<td>.861</td>
<td>.188</td>
</tr>
<tr>
<td>EX 6</td>
<td>.819</td>
<td>.048</td>
</tr>
<tr>
<td>EX 8</td>
<td>.816</td>
<td>.034</td>
</tr>
<tr>
<td>EX 10</td>
<td>.815</td>
<td>.025</td>
</tr>
<tr>
<td>EX 2</td>
<td>.795</td>
<td>.138</td>
</tr>
<tr>
<td>EX 4</td>
<td>.793</td>
<td>.219</td>
</tr>
<tr>
<td>EX 7</td>
<td>.788</td>
<td>.000</td>
</tr>
<tr>
<td>EX 5</td>
<td>.765</td>
<td>.134</td>
</tr>
<tr>
<td>EX11</td>
<td>.785</td>
<td>.009</td>
</tr>
<tr>
<td>EX 9</td>
<td>.783</td>
<td>-.013</td>
</tr>
<tr>
<td>EX 1</td>
<td>.770</td>
<td>.058</td>
</tr>
<tr>
<td>EX19</td>
<td>.036</td>
<td>.761</td>
</tr>
<tr>
<td>EX20</td>
<td>.069</td>
<td>.744</td>
</tr>
<tr>
<td>EX17</td>
<td>.073</td>
<td>.744</td>
</tr>
<tr>
<td>EX18</td>
<td>.084</td>
<td>.735</td>
</tr>
<tr>
<td>EX21</td>
<td>.088</td>
<td>.704</td>
</tr>
<tr>
<td>EX16</td>
<td>.029</td>
<td>.685</td>
</tr>
<tr>
<td>EX15</td>
<td>.075</td>
<td>.671</td>
</tr>
<tr>
<td>EX12</td>
<td>.057</td>
<td>.627</td>
</tr>
<tr>
<td>EX13</td>
<td>.086</td>
<td>.614</td>
</tr>
<tr>
<td>EX14</td>
<td>.086</td>
<td>.525</td>
</tr>
</tbody>
</table>

Specifically, Items EX1—EX11 loaded highly on the first factor. As these items were intended to measure students’ self-efficacy in taking the CET-4, this factor was
referred to as *self-efficacy* (SEE). Items EX12—EX21, which were designed to measure perceptions of the CET difficulty, loaded on the second factor. Thus this factor was labeled *task difficulty* (TAD).

Table 5.9 summarizes the descriptive statistics of the two factors concerning student expectations of taking the CET-4 test, including the questionnaire items in each of the factors as well as the mean, standard deviation, and skewness and kurtosis values of each factor. The means of SEE and TAD were 5.30 and 4.64 respectively, indicating that the students’ self-efficacy in taking the CET-4 was relatively high and their perceptions of the CET-4 task difficulty were relatively low, as also shown by the descriptive statistics in Table 5.7. The skewness and kurtosis values were close to 0, indicating that the data are normally distributed.

Table 5.9

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self efficacy (SEE)</td>
<td>EX1, EX2, EX3, EX4, EX5, EX6, EX7, EX8, EX9, EX10, EX11</td>
<td>5.30</td>
<td>1.25</td>
<td>-0.28</td>
<td>-0.03</td>
</tr>
<tr>
<td>Task difficulty (TAD)</td>
<td>EX12, EX13, EX14, EX15, EX16, EX17, EX18, EX19, EX20, EX21</td>
<td>4.64</td>
<td>1.28</td>
<td>-0.17</td>
<td>0.06</td>
</tr>
</tbody>
</table>

5.1.4 Test Preparation Practices

The fourth section of the questionnaire included 39 items that measured students’ CET-4 preparation practices. These items were intended for five sub-scales: 1) test preparation management; 2) cramming; 3) rehearsal test-taking strategies; 4) long-term skill development; and 5) social-affective strategies. Table 5.10 presents the descriptive statistics of these items.
Table 5.10

Descriptive Statistics of CET-4 Test Preparation

<table>
<thead>
<tr>
<th>Item</th>
<th>$M$</th>
<th>$SD$</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP1</td>
<td>5.43</td>
<td>1.62</td>
<td>-0.43</td>
<td>-0.69</td>
</tr>
<tr>
<td>TP2</td>
<td>5.35</td>
<td>1.55</td>
<td>-0.35</td>
<td>-0.66</td>
</tr>
<tr>
<td>TP3</td>
<td>5.10</td>
<td>1.55</td>
<td>-0.42</td>
<td>-0.59</td>
</tr>
<tr>
<td>TP4</td>
<td>5.31</td>
<td>1.56</td>
<td>-0.34</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP5</td>
<td>5.27</td>
<td>1.54</td>
<td>-0.30</td>
<td>-0.64</td>
</tr>
<tr>
<td>TP6</td>
<td>5.06</td>
<td>1.65</td>
<td>-0.24</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP7</td>
<td>5.44</td>
<td>1.45</td>
<td>-0.32</td>
<td>-0.60</td>
</tr>
<tr>
<td>TP8</td>
<td>5.67</td>
<td>1.35</td>
<td>-0.43</td>
<td>-0.62</td>
</tr>
<tr>
<td>TP9</td>
<td>5.46</td>
<td>1.48</td>
<td>-0.35</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP10</td>
<td>5.89</td>
<td>1.32</td>
<td>-0.57</td>
<td>-0.61</td>
</tr>
<tr>
<td>TP11</td>
<td>5.64</td>
<td>1.50</td>
<td>-0.48</td>
<td>-0.66</td>
</tr>
<tr>
<td>TP12</td>
<td>5.29</td>
<td>1.55</td>
<td>-0.30</td>
<td>-0.64</td>
</tr>
<tr>
<td>TP13</td>
<td>5.37</td>
<td>1.54</td>
<td>-0.36</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP14</td>
<td>5.13</td>
<td>1.60</td>
<td>-0.26</td>
<td>-0.61</td>
</tr>
<tr>
<td>TP15</td>
<td>5.12</td>
<td>1.49</td>
<td>-0.22</td>
<td>-0.56</td>
</tr>
<tr>
<td>TP16</td>
<td>5.23</td>
<td>1.58</td>
<td>-0.31</td>
<td>-0.66</td>
</tr>
<tr>
<td>TP17</td>
<td>5.19</td>
<td>1.57</td>
<td>-0.27</td>
<td>-0.61</td>
</tr>
<tr>
<td>TP18</td>
<td>5.63</td>
<td>1.52</td>
<td>-0.50</td>
<td>-0.67</td>
</tr>
<tr>
<td>TP19</td>
<td>6.04</td>
<td>1.36</td>
<td>-0.81</td>
<td>-0.45</td>
</tr>
<tr>
<td>TP20</td>
<td>5.94</td>
<td>1.29</td>
<td>-0.62</td>
<td>-0.58</td>
</tr>
<tr>
<td>TP21</td>
<td>5.67</td>
<td>1.45</td>
<td>-0.46</td>
<td>-0.69</td>
</tr>
<tr>
<td>TP22</td>
<td>5.06</td>
<td>1.57</td>
<td>-0.22</td>
<td>-0.59</td>
</tr>
<tr>
<td>TP23</td>
<td>6.07</td>
<td>1.30</td>
<td>-0.81</td>
<td>-0.40</td>
</tr>
<tr>
<td>TP24</td>
<td>5.63</td>
<td>1.40</td>
<td>-0.42</td>
<td>-0.61</td>
</tr>
<tr>
<td>TP25</td>
<td>4.89</td>
<td>1.78</td>
<td>-0.22</td>
<td>-0.71</td>
</tr>
<tr>
<td>TP26</td>
<td>4.55</td>
<td>1.79</td>
<td>-0.12</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP27</td>
<td>4.48</td>
<td>1.80</td>
<td>-0.10</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP28</td>
<td>4.93</td>
<td>1.76</td>
<td>-0.21</td>
<td>-0.70</td>
</tr>
<tr>
<td>TP29</td>
<td>3.94</td>
<td>1.89</td>
<td>0.03</td>
<td>-0.75</td>
</tr>
<tr>
<td>TP30</td>
<td>4.17</td>
<td>1.99</td>
<td>-0.02</td>
<td>-0.82</td>
</tr>
<tr>
<td>TP31</td>
<td>4.28</td>
<td>1.95</td>
<td>-0.07</td>
<td>-0.80</td>
</tr>
<tr>
<td>TP32</td>
<td>5.75</td>
<td>1.45</td>
<td>-0.57</td>
<td>-0.62</td>
</tr>
<tr>
<td>TP33</td>
<td>5.47</td>
<td>1.53</td>
<td>-0.39</td>
<td>-0.67</td>
</tr>
<tr>
<td>TP34</td>
<td>5.71</td>
<td>1.41</td>
<td>-0.49</td>
<td>-0.67</td>
</tr>
<tr>
<td>TP35</td>
<td>5.03</td>
<td>1.72</td>
<td>-0.28</td>
<td>-0.71</td>
</tr>
<tr>
<td>TP36</td>
<td>5.25</td>
<td>1.61</td>
<td>-0.31</td>
<td>-0.65</td>
</tr>
<tr>
<td>TP37</td>
<td>5.26</td>
<td>1.69</td>
<td>-0.34</td>
<td>-0.74</td>
</tr>
<tr>
<td>TP38</td>
<td>5.41</td>
<td>1.69</td>
<td>-0.43</td>
<td>-0.75</td>
</tr>
<tr>
<td>TP39</td>
<td>5.75</td>
<td>1.47</td>
<td>-0.56</td>
<td>-0.63</td>
</tr>
</tbody>
</table>

There were nine items (items TP25—TP31) with a mean score lower than 5.00, the smallest being 3.94 (item TP29: I keep on writing diaries/blogs in English.). These items were intended for the sub-scale of long-term English language knowledge and
skills development. Two items had a mean score over 6.00 (TP19, $M$=6.04: I go over the options beforehand so as to focus my attention accordingly in listening; TP23, $M$=6.07: While practicing writing, I try to avoid grammar and spelling mistakes). Both items concerned rehearsal of strategies related to CET-4 test taking. Thus students’ CET-4 preparation practices may have focused on practical aspects of test taking in order to obtain higher scores on the test. All the other items had a mean score between 5.00 and 6.00. The standard deviations of the items ranged from 1.29 (TP20) to 1.99 (TP30). The skewness and kurtosis values of all the items except the skewness of TP29 were negative, but none was smaller than -1.00, indicating a relatively normal distribution of the data.

The KMO measure of sampling adequacy for items in this section was .96. Bartlett’s test of sphericity was significant (p<.001). The PAF resulted in four factors. One complex variable (TP8) and three outlying variables (TP16, TP17 and TP22) were identified and deleted. After the deletion, the remaining variables were re-analyzed.

Items intended for the second sub-scale (cramming target knowledge & skills) and the third sub-scale (rehearsal test-taking strategies) loaded highly on the first factor. This factor was referred to as rehearsal and cramming (RAC). Items that loaded highly on the second factor were intended for use of long-term learning strategies to develop listening, reading, speaking and writing abilities in English. This factor was defined as long-term skill development (LSD). The third factor was referred to as test preparation management (TPM) as items that loaded highly on it were intended for this sub scale. Items that loaded highly on the fourth factor were intended for the sub scale of social-affective strategies (SAS). This factor was referred to as SAS. Table 5.11 presents the rotated factor matrix from the final PAF.
The long-term skill development strategy is essentially language learning through extensive exposure to and use of the target language. Both the test preparation management strategy and the rehearsal and cramming strategy are effort intensive but more directly related to the particular test. Using the test preparation management strategy, students indicated that they strategically used their time and resources during test preparation based on their evaluation of the test. Using the rehearsal and cramming strategy.
strategy, students focused their efforts on rehearsing and cramming test-taking skills. The social-affective strategy aims at enhancing performance indirectly through enhancing motivation and confidence. Descriptive statistics of the four factors and the items loading highly on each of them are presented in Table 5.12.

Table 5.12

Descriptive Statistics of Factors on CET-4 Test Preparation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rehearsal and cramming (RAC)</td>
<td>TP11, TP12, TP13, TP14, TP18,</td>
<td>5.33</td>
<td>1.16</td>
<td>-0.27</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>TP19, TP20, TP21, TP23, TP24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-term skill development (LSD)</td>
<td>TP25, TP26, TP27, TP28, TP29,</td>
<td>4.46</td>
<td>1.54</td>
<td>-0.07</td>
<td>-0.52</td>
</tr>
<tr>
<td></td>
<td>TP30, TP31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test preparation management (TPM)</td>
<td>TP1, TP2, TP3, TP4, TP5, TP6,</td>
<td>5.39</td>
<td>1.15</td>
<td>-0.28</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td>TP7, TP9, TP10, TP15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social-affective strategies (SAS)</td>
<td>TP32, TP33, TP34, TP35, TP36,</td>
<td>5.45</td>
<td>1.25</td>
<td>-0.36</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td>TP37, TP38, TP39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Of the four factors, LSD had the lowest mean score (4.46) and the highest standard deviation (1.54). SAS had the highest mean score (5.45), with a standard deviation of 1.25. These statistics suggested that social-affective strategies might be more used than were long-term language skill development strategies in the students’ CET-4 preparation. The other two factors (RAC and TPM) had a relatively high mean score and a low standard deviation. The skewness and kurtosis values, which were all negative, were close to 0, indicating a relatively normal distribution of the data.

5.2 Test Performance

A CET-4 test paper was used to collect data about students’ learning outcomes. The structure of the test paper is presented in Table 1.1. The sample of the CET-4 test
paper used in this study is included in Appendix G. Table 5.13 summarizes the
descriptive statistics of the section-level test score data.

Table 5.13

Descriptive Statistics of the CET-4 Section-level Test Score Data

<table>
<thead>
<tr>
<th>Section</th>
<th>Full score</th>
<th>Min.</th>
<th>Max.</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing (WT)</td>
<td>15</td>
<td>0</td>
<td>14</td>
<td>7.67</td>
<td>2.00</td>
<td>- .34</td>
<td>.89</td>
</tr>
<tr>
<td>Skimming and Scanning (RC)</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>7.40</td>
<td>2.08</td>
<td>-1.14</td>
<td>1.29</td>
</tr>
<tr>
<td>Short Conversations (LC)</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>4.50</td>
<td>2.10</td>
<td>- .06</td>
<td>- .96</td>
</tr>
<tr>
<td>Long Conversations (LC)</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>4.29</td>
<td>1.81</td>
<td>- .38</td>
<td>-.57</td>
</tr>
<tr>
<td>Listening Passages (LC)</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>4.70</td>
<td>2.27</td>
<td>-.29</td>
<td>-.21</td>
</tr>
<tr>
<td>Compound Dictation (LC)</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>4.46</td>
<td>2.41</td>
<td>.19</td>
<td>-.92</td>
</tr>
<tr>
<td>Banked Cloze (RC)</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>2.27</td>
<td>1.33</td>
<td>.19</td>
<td>-.60</td>
</tr>
<tr>
<td>Passage Reading (RC)</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>11.45</td>
<td>4.70</td>
<td>-.31</td>
<td>-.77</td>
</tr>
<tr>
<td>Cloze (IS)</td>
<td>10</td>
<td>0</td>
<td>9.5</td>
<td>5.05</td>
<td>1.94</td>
<td>-.37</td>
<td>-.28</td>
</tr>
<tr>
<td>Translation (WT)</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>2.51</td>
<td>1.24</td>
<td>-.34</td>
<td>-.76</td>
</tr>
</tbody>
</table>

The CET-4 was intended to measure four sub-constructs of the students’ English
language proficiency: listening comprehension (LC), reading comprehension (RC),
integrated skills (IS), and writing and translation (WT). In Table 5.13, the sub-construct
measured by each of the sections is indicated in brackets. In general, the students’ scores
on the skimming and scanning section were comparatively higher (7.40 out of 10) while
their scores on the compound dictation section were relatively lower (4.46 out of 10). The
absolute values of skewness and kurtosis, except the skewness and kurtosis of the score
on skimming and scanning, were all smaller than one, indicating a relatively normal
distribution of the data.

However, the PAF on the test score data resulted in three factors\(^7\). Table 5.14
presents the factor matrix with the loadings of the section score on each of the factors.

---

\(^7\) The CET score reporting system has recently been revised. Currently three sub-scores are reported (Listening Comprehension, Reading Comprehension and Writing and Translation). It is interesting to note that the PAF result reported in this study is consistent with the most updated CET score reporting.
The four sections that were intended to measure the sub-construct of listening comprehension (long conversation, short conversation, listening passages and compound dictation) loaded highly on the first factor. Thus this factor was referred to as listening comprehension (LIC). The cloze section, which was intended for a separate sub-construct, loaded highly on the second factor, together with the sections of passage reading and banked cloze. This factor was referred to as reading comprehension (REC). The skimming and scanning section, which was intended for reading comprehension, however, loaded highly on the third factor, together with the translation and writing section. Thus this factor was referred to as fast reading, writing, and translation (FWT).

Table 5.14

Rotated Factor Matrix of Test Score Data

<table>
<thead>
<tr>
<th>Section</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short conversation</td>
<td>.702</td>
<td>.211</td>
<td>.233</td>
</tr>
<tr>
<td>Long conversation</td>
<td>.667</td>
<td>.156</td>
<td>.204</td>
</tr>
<tr>
<td>Listening passages</td>
<td>.514</td>
<td>.215</td>
<td>.264</td>
</tr>
<tr>
<td>Compound dictation</td>
<td>.481</td>
<td>.340</td>
<td>.378</td>
</tr>
<tr>
<td>Cloze</td>
<td>.199</td>
<td>.793</td>
<td>.197</td>
</tr>
<tr>
<td>Passage reading</td>
<td>.229</td>
<td>.520</td>
<td>.285</td>
</tr>
<tr>
<td>Banked cloze</td>
<td>.409</td>
<td>.431</td>
<td>.229</td>
</tr>
<tr>
<td>Writing</td>
<td>.217</td>
<td>.159</td>
<td>.686</td>
</tr>
<tr>
<td>Translation</td>
<td>.303</td>
<td>.316</td>
<td>.503</td>
</tr>
<tr>
<td>Skimming and scanning</td>
<td>.215</td>
<td>.204</td>
<td>.455</td>
</tr>
</tbody>
</table>

Test scores for sections within each of the three factors were summed up to obtain the score for each of the factors. Table 5.15 summarizes the descriptive statistics for the three factors (listening comprehension, reading comprehension, and fast reading, writing and translation), including the sections included in each factor, possible score range, mean, standard deviation, skewness and kurtosis. The summative factor scores were used in the SEM analysis as construct indicators for the test performance.
Table 5.15

Descriptive Statistics for Learning Outcomes

<table>
<thead>
<tr>
<th>Factors</th>
<th>Sections included</th>
<th>Full score</th>
<th>Range</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening comprehension (LIC)</td>
<td>Short conversations</td>
<td>35</td>
<td>0-35</td>
<td>17.95</td>
<td>6.67</td>
<td>0.01</td>
<td>-.70</td>
</tr>
<tr>
<td></td>
<td>Long conversations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Listening passages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compound dictation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading comprehension (REC)</td>
<td>Banked cloze</td>
<td>35</td>
<td>1-33.5</td>
<td>18.77</td>
<td>6.64</td>
<td>-0.30</td>
<td>-0.65</td>
</tr>
<tr>
<td></td>
<td>Passage reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cloze</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fast reading, writing and translation (FWT)</td>
<td>Translation</td>
<td>30</td>
<td>1.5-28</td>
<td>17.57</td>
<td>4.16</td>
<td>-0.80</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skimming and scanning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Related to statistics in Table 5.13, statistics in Table 5.15 showed that the mean score on FWT was comparatively higher (17.57 out of 30) and its standard deviation was relatively smaller (4.16) while the mean score on LIC was comparatively lower (17.95 out of 35) and its standard deviation was relatively bigger (6.67). These statistics suggested that the students’ performances on FWT were comparatively higher and there was relatively small variation in their performances while their performances on LIC were comparatively lower on average and there was relatively big variation in their performances.

5.3 Relationships Among Perceptions of CET-4, Test-taking Expectations, Test Preparation Practices, and Test Performance

In order to investigate the relationships among students’ perceptions of the CET-4 demands and uses as well as their test-taking expectations, test preparation practices, and test performance, a CET-4 washback model was proposed in this study (see Figure 5.1). To assess this hypothesized model, SEM was performed using the results reported in
Sections 5.1 and 5.2, particularly results from the PAFs. By evaluating the washback model using empirical data in this study, RQ3 was addressed.

### 5.3.1 The CET-4 Washback Model

There are five latent variables in the proposed model: perceptions of test demands (Pdemand), perceptions of test uses (Puse), test-taking expectations (Expect), test preparation practices (Prep), and test performance (Perform). Results from the PAFs on the questionnaire data and the test score data were used in the SEM as construct indicators for these latent variables. Table 5.16 summarizes these latent variables and their indicators.

#### Table 5.16

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Construct indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptions of test demands</td>
<td>Perceptions of English language knowledge and skills (LKS)</td>
</tr>
<tr>
<td></td>
<td>Perceptions of listening skills (LIS)</td>
</tr>
<tr>
<td></td>
<td>Perceptions of background knowledge (BAK)</td>
</tr>
<tr>
<td></td>
<td>Perceptions of reading skills (RES)</td>
</tr>
<tr>
<td>Perceptions of test uses (Puse)</td>
<td>Task values (TAV)</td>
</tr>
<tr>
<td></td>
<td>Instrumental uses (INU)</td>
</tr>
<tr>
<td></td>
<td>Achievement uses (ACU)</td>
</tr>
<tr>
<td>Test-taking expectations (Expect)</td>
<td>Self-efficacy (SEE)</td>
</tr>
<tr>
<td></td>
<td>Perceptions of task difficulty (TAD)</td>
</tr>
<tr>
<td>Test preparation practices (Prep)</td>
<td>Rehearsal and cramming (RAC)</td>
</tr>
<tr>
<td></td>
<td>Long-term skill development (LSD)</td>
</tr>
<tr>
<td></td>
<td>Test preparation management (TPM)</td>
</tr>
<tr>
<td></td>
<td>Social-affective strategies (SAS)</td>
</tr>
<tr>
<td>Test performance (Perform)</td>
<td>Listening comprehension (LIC)</td>
</tr>
<tr>
<td></td>
<td>Reading comprehension (REC)</td>
</tr>
<tr>
<td></td>
<td>Fast reading, writing and translation (FWT)</td>
</tr>
</tbody>
</table>

As shown in Table 5.16, the five latent variables in the CET-4 washback model were measured by 16 indicators based on the PAF results. Descriptive statistics of these indicators, including means, standard deviations, skewness, and kurtosis, are presented in
Tables 5.3, 5.6, 5.9, 5.12, and 5.15 respectively. As reported in these tables, the skewness and kurtosis values indicated normality of the distributions of the indicators. Histograms of all indicators were examined and revealed that univariate normality could be held in them. Table 5.17 presents the correlation matrix for the indicators.

Table 5.17

Correlation Matrix for Construct Indicators in the CET-4 Washback Model

<table>
<thead>
<tr>
<th></th>
<th>LKS</th>
<th>LIS</th>
<th>BAK</th>
<th>RES</th>
<th>TAV</th>
<th>INU</th>
<th>ACU</th>
<th>SEE</th>
<th>TAD</th>
<th>LSD</th>
<th>RAC</th>
<th>TPM</th>
<th>SAS</th>
<th>LIC</th>
<th>REC</th>
<th>FWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>LKS</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIS</td>
<td>.62</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAK</td>
<td>.55</td>
<td>.60</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RES</td>
<td>.67</td>
<td>.62</td>
<td>.58</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAV</td>
<td>.36</td>
<td>.34</td>
<td>.42</td>
<td>.34</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INU</td>
<td>.57</td>
<td>.53</td>
<td>.45</td>
<td>.53</td>
<td>.55</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACU</td>
<td>.56</td>
<td>.50</td>
<td>.42</td>
<td>.50</td>
<td>.63</td>
<td>.67</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEE</td>
<td>.53</td>
<td>.49</td>
<td>.38</td>
<td>.46</td>
<td>.41</td>
<td>.50</td>
<td>.55</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAD</td>
<td>.29</td>
<td>.29</td>
<td>.39</td>
<td>.27</td>
<td>.30</td>
<td>.40</td>
<td>.30</td>
<td>.19</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSD</td>
<td>.23</td>
<td>.28</td>
<td>.32</td>
<td>.25</td>
<td>.39</td>
<td>.39</td>
<td>.40</td>
<td>.39</td>
<td>.27</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAC</td>
<td>.55</td>
<td>.56</td>
<td>.45</td>
<td>.53</td>
<td>.50</td>
<td>.57</td>
<td>.63</td>
<td>.58</td>
<td>.29</td>
<td>.49</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPM</td>
<td>.54</td>
<td>.56</td>
<td>.45</td>
<td>.52</td>
<td>.46</td>
<td>.55</td>
<td>.58</td>
<td>.54</td>
<td>.31</td>
<td>.53</td>
<td>.63</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAS</td>
<td>.50</td>
<td>.51</td>
<td>.39</td>
<td>.46</td>
<td>.45</td>
<td>.53</td>
<td>.62</td>
<td>.55</td>
<td>.27</td>
<td>.55</td>
<td>.69</td>
<td>.60</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIC</td>
<td>-.01</td>
<td>.00</td>
<td>-.05</td>
<td>-.07</td>
<td>.00</td>
<td>-.02</td>
<td>.07</td>
<td>.06</td>
<td>-.12</td>
<td>.01</td>
<td>.00</td>
<td>-.07</td>
<td>.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>.09</td>
<td>.09</td>
<td>.04</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.11</td>
<td>.11</td>
<td>-.04</td>
<td>.08</td>
<td>.08</td>
<td>.03</td>
<td>.07</td>
<td>.55</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>FWT</td>
<td>.05</td>
<td>.07</td>
<td>-.02</td>
<td>.03</td>
<td>.10</td>
<td>.01</td>
<td>.09</td>
<td>.10</td>
<td>-.07</td>
<td>.00</td>
<td>.05</td>
<td>.04</td>
<td>.06</td>
<td>.54</td>
<td>.50</td>
<td>1.00</td>
</tr>
</tbody>
</table>

No violation of linearity was found by examining scatter plots of all possible pairs of the variables. Pairwise multicollinearity was checked by inspecting the correlation matrix of the variables. Correlations above .85 generally suggest a problem (Kline, 2005). Based on this criterion, no extremely high value of correlation coefficient was found in Table 5.17, the highest being .69 (correlation of RAC and SAS).

The covariance matrix underlying the 16 indicators is presented in Appendix J. A structural equation model was used to test the fit of this matrix to the proposed CET-4
washback model. Figure 5.2 presents the structural equation model generated by LISREL.

In Figure 5.2, the 16 construct indicators are represented by the rectangles and the five latent variables are represented by the ellipses. There is a residual term attached to each indicator, representing the amount of variation that is due to measurement error. There is also an error term associated with each of the latent variables representing that part of the latent variable not accounted for by the linear influences of the other variables specified in the model.

The latent variables are connected in the model in order to reflect the relationships between students’ perceptions of the CET-4 demands and uses, their CET-4 test-taking expectations, test preparation practices, and test performance. The relationships between

Figure 5-2. CET-4 washback model generated by LISREL.
the latent variables, and the relationships between the latent variables and the indicators are the main focus of this study. These relationships are presented graphically in the diagram by arrows. The one-way arrows indicate that the variable at the end of the arrow is explained by the variable at the beginning of the arrow. For example, the arrows connecting the five latent variables with the indicators denote that the latent variables account for the variations in the indicators. In a similar vein, the arrows connecting the five latent variables indicate that students’ perceptions of the CET-4 demands and uses as well as their test-taking expectations are hypothesized to explain their test preparation practices, which in turn is assumed to explain their test performance. However, causal relations among the latent variables are not assumed in this study considering the research design. In Figure 5-2, the two-way arrows are used to designate covariation between the three latent variables (perceptions of test demands, perceptions of test uses, and test-taking expectations).

In order to make the decision whether the CET-4 washback model could be tentatively relied upon or should reasonably be rejected as a means of explaining the covariance matrix data, the adequacy and appropriateness of the model were evaluated based on two criteria: 1) values of selected global model fit indices; and 2) individual parameter estimates.

5.3.2 Evaluation of the Global Fit of the CET-4 Washback Model

The LISREL program reported several goodness-of-fit indices. It is generally recommended that multiple indices be considered simultaneously when a global model fit is evaluated (Lei & Wu, 2007). The selection of model fit indices in this study was based on the 2-index strategy proposed by Hu and Bentler (1999) (i.e., reporting SRMR along
with one of the fit indices such as CFI, AGFI, and RMSEA). Hu and Bentler (1999) also suggested the following criteria for an indication of good model-data fit using these indices: SRMR≤.08, CFI≥.95, AGFI≥.90, and RMSEA ≤.06. In addition, the chi-square index (χ²) was also reported in this study as it is the traditional measure for evaluating overall model fit. A good model fit would provide an insignificant χ² result at a 0.05 threshold. It should be noted, however, there exist a number of limitations in using χ² as a fit statistic in SEM because it is affected by factors such as sample size, model size, and distribution of variables. Due to the restrictiveness of χ², alternative indices have been suggested. For example, Wheaton, Muthen, Alwin, and Summers (1977) developed the relative/normed chi-square test statistic (χ²/df (degree of freedom) ratio). There is a lack of general consensus with regards to the acceptable ratio for this statistic with recommendations ranging from 5.0 (Wheaton et al., 1977) to a more conservative estimate of 2.0 (Tabachnick & Fidell, 2012). Table 5.18 summarizes values of the selected global model fit indices in this study.

Table 5.18

<table>
<thead>
<tr>
<th>SRMR</th>
<th>CFI</th>
<th>AGFI</th>
<th>RMSEA</th>
<th>χ²</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>.046</td>
<td>.98</td>
<td>.91</td>
<td>.059</td>
<td>236.89</td>
<td>97</td>
</tr>
</tbody>
</table>

Statistics in Table 5.18 indicated that the selected indices were all satisfactory except that the χ² value was statistically significant (p=.00). The χ² /df ratio (2.44) was low. Therefore, at the global level, the CET-4 washback model presented in Figure 5-1 and Figure 5-2 demonstrated a reasonable fit to the data collected in this study.
5.3.3 Parameter Estimation for the CET-4 Washback Model

After the global fit of the model was evaluated, individual parameter estimates were examined. Table 5.19 provides the standardized (also presented in Figure 5-2) and unstandardized parameter estimates, and their standardized error estimates as well as the \(p\)-values.

Table 5.19

Parameter and Standard Error Estimates for the CET-4 Washback Model

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Standardized Estimate</th>
<th>Unstandardized Estimate</th>
<th>Standard Error</th>
<th>(P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loadings/effects on Pdemand</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LKS</td>
<td>.81</td>
<td>1.00(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIS</td>
<td>.79</td>
<td>1.01</td>
<td>.06</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>BAK</td>
<td>.71</td>
<td>1.06</td>
<td>.07</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>RES</td>
<td>.80</td>
<td>1.08</td>
<td>.06</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Loadings/effects on Puse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAV</td>
<td>.70</td>
<td>1.00(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INU</td>
<td>.80</td>
<td>.95</td>
<td>.06</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>ACU</td>
<td>.86</td>
<td>1.12</td>
<td>.07</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Loadings/effects on Expect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEE</td>
<td>.57</td>
<td>1.00(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAD</td>
<td>.34</td>
<td>.61</td>
<td>.08</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Loadings/effects on Prep</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSD</td>
<td>.61</td>
<td>1.00(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAC</td>
<td>.84</td>
<td>1.04</td>
<td>.08</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>TPM</td>
<td>.78</td>
<td>.96</td>
<td>.08</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>SAS</td>
<td>.81</td>
<td>1.07</td>
<td>.08</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Loadings/effects on Perform</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIC</td>
<td>.76</td>
<td>1.00(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>.72</td>
<td>.98</td>
<td>.09</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>FWT</td>
<td>.71</td>
<td>.57</td>
<td>.05</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Path Coefficients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pdemand→Prep</td>
<td>.52</td>
<td>.58</td>
<td>.20</td>
<td>.003</td>
</tr>
<tr>
<td>Puse→Prep</td>
<td>.92</td>
<td>.82</td>
<td>.18</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Expect→Prep</td>
<td>-.41</td>
<td>-.53</td>
<td>.34</td>
<td>NS(^b)</td>
</tr>
<tr>
<td>Prep→Perform</td>
<td>.06</td>
<td>.31</td>
<td>.31</td>
<td>NS(^b)</td>
</tr>
<tr>
<td>Covariances</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pdemand and Puse</td>
<td>.76</td>
<td>.66</td>
<td>.07</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Pdemand and Expect</td>
<td>1.07</td>
<td>.64</td>
<td>.07</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Puse and Expect</td>
<td>1.11</td>
<td>.84</td>
<td>.09</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Note. Table values are maximum likelihood estimates. \(^a\)Fixed parameter to set the scale of the latent variable  \(^b\)Non significant.
The model solution was considered proper because there were no out-of-range parameter estimates and standard error estimates were of similar magnitude. All parameter estimates, except the path coefficient values for the paths from Expect to Prep and from Prep to Perform, were considered large (not likely zero) at the critical level of .05. Standardized factor loadings in measurement models should fall between 0 and 1 with higher values suggesting better indicators for the latent variable (Lei & Wu, 2007). Standardized loadings presented in Table 5.19 indicated that the indicators were satisfactory for the latent variables they loaded on. “Coefficients for the structural paths are interpreted in the same way as regression coefficients” (Lei & Wu, 2007, p. 37). The standardized coefficient value of 0.52 for the path from Pdemand to Prep suggested that as students’ perceptions of the CET-4 test demands increased by one standard deviation, their test preparation practice was expected to increase by 0.52 of a standard deviation. The standardized value of 0.92 for the path from Puse to Prep revealed that for every standard deviation increase in students’ perceptions of the CET-4 test uses, their test preparation practices were expected to increase by 0.92 of a standard deviation.

Overall, the SEM results reported above suggested that the CET-4 washback model proposed in this study could be accepted to present relationships among students’ perceptions of the test, their test-taking expectations, test preparation practices, and test performance. These results also indicated that both the perceptions of the CET-4 demands and the perceptions of the CET-4 uses had significant influences on test preparation. Comparatively, the effect of perceptions of test uses was greater than that of perceptions of test demands, indicating the educational and social impact of the use of CET-4 within the Chinese tertiary EFL context.
Chapter 6 Discussion and Conclusions

Within the argument-based validation framework (Kane, 2006, 2013), this chapter discusses major results reported in Chapters 4 and 5 and draws conclusions based on the findings to address the following three research questions of this study:

RQ1. What are the intended consequences of the CET-4 and what are the major procedures taken by the test developer to achieve the intended consequences?

RQ2. What are the actual uses of the CET-4 in the educational, business, and government contexts and what are the value implications underlying the actual uses?

RQ3. How does the CET-4 influence students’ learning and learning outcomes?

By answering these questions, this study addresses the research problem why the CET-4 washback exists in the Chinese tertiary educational context.

Findings of this study are discussed within the IUA framework for the CET-4 presented by Figure 2-2, whereby the policy decision of using the CET-4 program for educational reform is evaluated in terms of the utilization inference and the decision rule. The utilization inference is based on all the other inferences in the IUA. Findings about the CET-4 developer’s efforts to achieve the intended consequences provide evidence pertinent to these inferences. The decision rule is evaluated in terms of intended and unintended consequences. Findings about the CET-4 washback and the actual uses of the test in the educational and societal contexts constitute evidence regarding consequences of implementing the testing program.

Table 6.1 relates the research questions as well as major findings of this study to the inferences and the decision rule in the IUA.
Table 6.1

Research Questions and Findings in Relation to the Inferences in the IUA

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant</th>
<th>Findings</th>
<th>RQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Domain</td>
<td>Criterion upon which test is designed identifies English knowledge and</td>
<td>CECR specified English language requirements</td>
<td>RQ1</td>
</tr>
<tr>
<td>Description</td>
<td>skills.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring</td>
<td>Scoring rule is appropriate and scoring is free of bias.</td>
<td>Quality control of CET-4 marking process</td>
<td>RQ1</td>
</tr>
<tr>
<td>Generalization</td>
<td>Observed score is generalizable over the relevant parallel versions of test.</td>
<td>Standardization of CET-4</td>
<td>RQ1</td>
</tr>
<tr>
<td>Explanation</td>
<td>Expected scores are attributed to a construct of English proficiency.</td>
<td>Score factor structure and construct label of CET-4</td>
<td>RQ1</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>Test construct accounts for quality of students’ performance in target domain.</td>
<td>CET-4 was aligned with curriculum. Test score was correlated to classroom performance</td>
<td>RQ1</td>
</tr>
<tr>
<td>Utility</td>
<td>The CET-score provides useful information for its legitimate uses in the educational context</td>
<td>Reform of the CET-4 score reporting system</td>
<td>RQ1</td>
</tr>
<tr>
<td>Decision Rule</td>
<td>Intended consequences are achieved with acceptable unintended consequences</td>
<td>Actual uses and washback of the CET-4</td>
<td>R2, R3</td>
</tr>
</tbody>
</table>

This chapter consists of seven sections. Section 6.1 discusses findings regarding intended consequences of the CET-4 from the test developer’s perspective to address first portion of RQ1. This discussion reveals the rationale underlying the implementation of the CET-4 testing program. Section 6.2 discusses findings concerning the test developer’s efforts to achieve the intended consequences. This discussion focuses on the CET-4 developer’s justifications for different inferences in the IUA to address second portion of RQ1. Section 6.3 discusses uses of the CET-4 score in the educational, business, and government contexts to address RQ2. This discussion highlights the multiple interpretations and multiple uses of the test and points to the value implications.
underlying the test. These test uses are discussed in relation to the intended consequences of the test discussed in Section 6.1. Section 6.4 focuses on the results about washback of the CET-4 (i.e., its effects on students’ learning and learning outcomes) to address RQ3. In order to understand what brings about this washback, these results are linked to findings concerning the test developer’s efforts to support the IUA discussed in Section 6.2, and findings regarding the actual uses of the CET-4 discussed in Section 6.3. As washback is considered a major dimension of the test consequences, the washback results are discussed, again, in relation to the intended consequences discussed in Section 6.1.

Following the above discussions, Section 6.5 makes concluding remarks drawing on the major findings of this study. Section 6.6 outlines the limitations of the study and points out directions for future research. Finally, Section 6.7 summarizes contributions of this study and presents implications of the findings for the CET developer and for College English teaching and learning.

### 6.1 Test Developer’s Intended Consequences of the CET-4

The CET-4 is an English language testing program initiated and implemented by the CMoE for all university undergraduates. It is a policy initiative, together with the implementation of the College English curriculum (i.e., the CECR), to promote tertiary EFL education in China. In a context in which testing programs are used for policy implementation, the distinction between the test developer and the test user is blurred (Chalhoub-Deville, 2009). The policy-maker assumes the roles of both the test developer and the test user. In the CET-4 context, the CMoE makes policies regarding test requirements and intended uses, and appoints the NCETC to be responsible for the
development of the test. Therefore, both the CMoE and the NCETC are regarded as the test developer in this study.

This study found two major intended consequences of the CET-4 from the test developer’s perspective: 1) to promote College English education in China; and 2) to provide useful information about students’ English language proficiency based on the curriculum.

6.1.1 Promoting College English Education in China

Analysis of the documents disseminated by the test developer shows that promoting EFL education in Chinese tertiary education has consistently been the goal of implementing the testing program. The test, coupled with the College English curriculum, is intended to enhance the quality of College English education. As mentioned by Deyi in the interview (see Section 4.2.2), both the revision of the College English curriculum and the reform of the CET-4 are major parts of the CERP, which, in turn, is part of the HEQRP initiated by the CMoE with the purpose of improving the overall quality of China’s tertiary education.

This intended consequence is consistent with the global trend of using large-scale testing programs as policy instruments to improve educational quality (e.g., Klenowski & Wyatt-Smith, 2012; Klinger et al., 2008; Levin, 1998; McEwen, 1995; Nagy, 2000; Roach & Frank, 2007; Rogers, 2014; Shulha & Wilson, 2009; Stevenson, 1996; Wagner, 2010). A rationale underlying this trend is that the curriculum makes explicit the expected outcomes for students’ achievement at specific points in educational programs, and a testing program is implemented to assess these outcomes. It is hoped that to the extent
that the test is aligned with the curriculum, the quality of education can be reported, monitored, accounted for, and improved (Cumming, 2009).

College English education has been given policy priority by the CMoE over the past three and half decades because of sociopolitical and educational reasons. Politically, the quality of higher education in general and the quality of EFL education in particular are considered by the government a key factor to the nation’s strength in the global knowledge economy in the 21st Century (CMoE, 2002; Mok, 2005). Socially, the exponential increase of the number of students enrolled in Chinese tertiary institutions since late 1990s has generated concerns over the quality of higher education in China. Educationally, College English is the only course required of all the undergraduate students across Chinese universities and colleges. There are growing concerns over the “low efficiency and effectiveness” of the College English education in developing students’ English use abilities (Dai, 2001; Jing, 1999; Yang, 2015; Zhao & Coniam, 2008).

In this context, the CET-4 plays a prominent role in the implementation of the EFL policy in China. The CMoE, as the central authority, granted English language special priority and status. This language policy is explicitly stated through official documents such as the College English curriculum and the CET-4 test syllabus. The CET-4 is a mechanism used by the CMoE to create, perpetuate, and manipulate this language policy. As noted by Shohamy (2007, p. 117), “the introduction of language tests in certain languages delivers messages and ideologies about the prestige, priorities and hierarchies of certain language(s)”. The central authority uses tests to control large-scale educational systems by defining what kind of knowledge is prestigious (Shohamy, 2001).
In the context where language tests are used as policy instruments, test constructs are authorized by policy (McNamara, 2010). Results of this study show how the definition of English proficiency changed in the College English curriculum with the changing needs in the society for university graduates with English use abilities. There was a shift of focus from English language knowledge and reading to a more broad definition of English use abilities, particularly listening and speaking. Accordingly, the CET-4 was revised in terms of its content and format. However, the broad policy-determined construct description presents methodological challenges for the NCETC in terms of test design, particularly within the practicality of the CET-4. I will further discuss these challenges in Section 6.2.

The intended social consequences of the CET-4 explain, in part, the multiple uses of the test in China. As Fulcher and Davidson (2007) noted, language tests used as policy instruments are usually co-opted into services for which they were not designed. The misuses and abuses of the CET-4 score have been reported in previous studies (e.g., Wang & Sun 2012) and become a fundamental concern of the test developer (Jin, 2011). Findings of this study about the actual uses of the test provide further evidence concerning the power of the test. Test users in this study sought to justify their test use decisions by acknowledging the status of the English language in China as well as the authority of the CET-4 developer. I will further discuss these findings in Section 6.3.

6.1.2 Informing College English Education

As mentioned above, promoting College English education in China has been a consistent goal of implementing the CET-4 over the past few decades. However, there have been changes with regard to the approaches to achieve this goal, with a shift of
focus from motivation to informing. Initially, all the non-English major undergraduate students in Chinese universities and colleges were required to pass the CET-4 to graduate and/or obtain their academic degrees. The purpose of this policy was to motivate students to study English. Currently, the CET-4 developer paid increasing attention to both intended and unintended consequences of the test and focused its efforts on enhancing the utility of the test score in terms of informing College English teaching and learning.

The CET-4 was intended to provide useful information to test users in the educational context about students’ English language proficiency based on the College English curriculum. This intended consequence was clearly stated in the current CET-4 test syllabus and was repeatedly emphasized by Deyi in the interview. Using the CET-4 results to inform teaching and learning was based on the assumption that information gathered for a summative purpose can also serve to help teaching and learning (Sambell, McDowell, & Montgomery, 2013). Aligned with the curriculum, a large-scale assessment can effectively inform classroom practices (EQAO, 2012a; Klinger et al, 2008; Pellegrino, 2014; Phelps, 2012; Porter & Smithson, 2000). The CET-4 was designed based on the College English curriculum. Specifically, the requirements for students at the basic level specified in the CECR (Table 4.3) defined the target domain of the CET-4. Based on this target domain, the CET-4 score was intended to provide useful information about students’ English proficiency related to the basic requirements specified in the CECR. Deyi’s explanations of the usefulness of the CET-4 results as well as results about how the CET-4 results were used in the two universities in this study provided evidence about the relevance and utility of the test score for College English teaching and learning.
There are examples showing that evidence collected from summative assessments can be used for formative purposes (Rogers, 2014). The usefulness of this information for improving teaching and learning would depend on factors such as the frequency of the test issued to test-takers, the details of the information the test could provide, as well as the pressure exerted by the external testing requirements (Harlen, 2006). Findings of this thesis study concerning the CET-4 developer’s efforts to achieve the intended consequences of the test indicate that the CET-4 was moving in this direction. I will discuss these efforts in the following section.

6.2 Test Developer’s Efforts to Achieve Intended Consequences

This section discusses measures taken by the CET-4 developer to achieve the intended consequences of the test. These efforts include technical procedures as well as policy strategies to enhance the usefulness of the CET-4 and provide evidence for the different inferences in the IUA.

6.2.1 Efforts in Support of the Decision Rule

The CET-4 developer made continuous efforts at the test design level and the policy level to promote the intended consequence of the test in terms of enhancing College English education and to mitigate its unintended impact in both the educational and societal contexts.

Washback is a major consideration of the CET-4 developer. At the test design level, the NCETC has continuously revised the CET-4 test format and content. This is consistent with Messick’s (1994, 1996) suggestion that sources of invalidity should be minimized in test design in order to reduce negative washback and to promote positive washback. A major measure is the decrease of the percentage of MCQs and the increase
of the percentage of constructed-response questions. The multiple-choice format was considered as a potential source of construct irrelevant variance and its negative influences on teaching and learning were discussed in literature in both educational assessment in general (e.g., Sato & Ikeda, 2015; Shepard, 1990) and in the CET-4 in particular (e.g., Gu, 2005). However, MCQs still accounted for 55%--70% of the CET-4 total score. This may be necessary given the large scale of the test.

Findings of this study and other studies (e.g., Hillocks, 2002; Li, Zhong, & Suen, 2012) indicate that essay writing might not be immune to narrow test preparation and test-taking strategies. Deyi pointed out that some students may prepare for the CET-4 writing by memorizing some model essays or lengthy chunks of text. This is consistent with Madaus’ (1988) observation that students tend to employ the same memorized formulae in response to different prompts in essay writing, as well as Linn et al.’s (1991) assertion that direct assessment will not necessarily foster positive classroom activities for learning. A direct test format such as writing might encourage the use of a formulaic approach that is effective in generating high-scoring essays within the time limit imposed by the test, but fails to develop deep learning. In this respect, the effectiveness of using improved test design to enhance intended positive washback may be limited.

The CET-4 developer’s concerns about consequences go beyond the measurement level to the policy level as the testing program is used as a policy instrument. Consequences are a particularly critical issue in the CET-4 context, given 1) its intended educational and sociopolitical purposes; 2) its large test-taker population; 3) the high stakes associated with its uses; and 4) the controversy over its uses, misuses and abuses in
the Chinese context, where large-scale high-stakes testing has been playing a powerful and yet socially accepted role.

Findings of this study indicate that unintended consequences of the CET-4 in both the educational and societal contexts received increased attention from the test developer. For example, an assessment framework was proposed in the current College English curriculum, which highlights the summative function of the CET-4 and emphasizes the importance of evaluating students’ English proficiency using evidence from multiple sources. The CET-4 certificate was replaced with a form reporting both the total score and the sub-scores for listening, reading, integrated skills, and translation and writing. These measures were taken for two purposes. First, the removal of the CET-4 certificate was intended to mitigate the power of the test by discouraging universities from using it as an indispensable requirement for graduation or degree conferment and employers from making it a prerequisite for employment. Second, it was expected that the reported CET-4 score, including the total score and the sub-scores, would provide more relevant information for test users in the educational context. This is consistent with Rogers’ (2014) suggestion that in large-scale assessment sub-scores should be reported given that the curriculum is multidimensional. This information could be used together with information about students’ English language proficiency from other sources for the purpose of improving teaching and learning. However, Chen’s (2011) study found that the assessment policy changes (i.e., the shift to optional status of the CET-4 and the incorporation of formative assessment) did not take place in universities. Chen (2011) asserts that sociocultural factors are decisive and responsible for the limited effect of the reform.
The CET-4 is implemented as a policy instrument to promote College English education in China. Therefore, consequences of the testing program are a major consideration of the CET-4 developer. Findings of this thesis study indicate that consequences of the CET-4 are a function of the test and the context in which it is implemented. Validity of the use of language assessment in the service of policy should be discussed in the broader educational, social, and political contexts in which it is commissioned (Chalhoub-Deville, 2009; Fulcher, 2009; McNamara, 2010; McNamara & Roever, 2006; Shohamy, 2007).

6.2.2 Efforts in Support of the Utilization Inference

In order to enhance the usefulness of the CET-4 score, the test developer focused its efforts on the quality of the test. The overall usefulness of a test is a function of qualities such as reliability, construct validity, authenticity, interactivity, and impact based on practicality of the test (Bachman & Palmer, 1996). Within the IUA framework, usefulness of the CET-4 is related to the utilization inference, which is supported by all the other inferences leading to it (see Figure 2-2). The following discussion focuses on the CET-4 developer’s efforts in support of the utilization inference in relation to these inferences.

The IUA for the CET-4 presented in Figure 2-2 starts with target domain description. The target domain of the CET-4 is the requirements for students at the basic level specified in the CECR (Table 4.3). These requirements include narrowly and precisely described language use skills and abilities in a limited range of contexts. This narrow and precise definition of the target domain is useful from the perspective of test design. It allows the CET-4 testing procedures to be focused on tasks in restricted
contexts requiring English language skills in listening, reading, writing, and translation as specified in the CECR. Deyi highlighted this usefulness in the interview by highlighting the connections between the curriculum, the test specifications, and the test design. Alignment between the curriculum and the test is also essential in terms of informing classroom practices (Klinger et al., 2008; Pellegrino, 2014).

Results reported in Section 4.2.1 provide evidence for the scoring inference in the CET-4 IUA. This inference assigns scores to students’ performances (observations) to obtain their observed scores on specific tasks of the CET-4 or the whole test. According to Kane (2006, p. 34), the scoring inference “relies on the assumption that the scoring criteria are appropriate and are applied as intended, that the process is free of bias, and that any statistical models (scaling and equating) employed in scoring are appropriate.” In the interview, Deyi emphasized the NCETC’s quality control procedures in the CET-4 marking process. Systematic procedures of quality control were established and effective measures were taken by the NCETC in terms of marker training and marking monitoring to ensure the inter- and intra- reliability of the scores assigned to the constructed-response questions. These efforts provide evidence for the assumptions underlying the scoring inference that scoring criteria for the constructed questions are applied as intended and that the marking process is free of bias associated with markers. Deyi also highlighted the CET-4 score equation process as a measure to ensure accurate interpretation of the CET-4 score as a measure of students’ English proficiency based on the curriculum.

Results reported in Section 4.2.1 also provide evidence for the generalization inference. This inference extends the interpretation of the test results from a claim about students’ observed scores on one test to a claim about their expected scores across similar
assessments. The primary measure taken by the CET-4 developer related to this inference is the standardization of the test procedures. The CET-4 is described as a standardized test on the CET official website and Deyi provided a detailed description of the standardization throughout the test administration process. The purpose of standardization is to control random error in the test process. It not only ensures that all test takers are asked to perform the same tasks under the same conditions but also makes students’ performances on different administrations of the CET-4 comparable. However, standardization of test procedures introduces systematic errors into estimates of the target score, particularly in the context where the construct of the test is broadly defined (Kane, 2006). Systematic errors constitute a source of construct irrelevant variance threatening test validity (Messick, 1996).

With regard to the explanation inference, the CET-4 construct is labeled English proficiency, which is broadly described and given an expansive interpretation in the College English curriculum and the CET-4 documents. The broad definition of the CET-4 construct is determined by the goal of College English education stated in the CERP, (i.e., to develop students’ English language proficiency to satisfy the social needs for university graduates with abilities to use English). To be consistent with this goal, the CET-4 construct is labeled English proficiency in the test and curriculum documents. Thus the wording of the test construct is determined by policy procedures. Test constructs are authorized by policy in the context where language tests are used as policy instruments (McNamara, 2010).

Nevertheless, the broad policy-determined construct description presents methodological challenges for the NCETC in terms of test design, particularly within the
practicality of the CET-4. To address these challenges, the CET-4 is designed based on the requirements for students at the basic level specified in the CECR, which narrowly describes language use skills and abilities in a limited range of contexts. Thus there is inconsistency between the CET-4 target domain and the construct label of the test. This inconsistency was highlighted by the two participants in the educational context. As they pointed out in the interviews (see Section 4.3.1), the CECR gave a broad and “ambitious definition” of English proficiency including everything from listening, speaking, reading, and writing to cross cultural communication competence, which is beyond the limited range of tasks included in the CET-4. The inconsistency between the restricted target domain of the CET-4 and the broad description of the test construct gives rise to construct under-representation, a major threat to test validity (Messick, 1989). Kane (2013) maintains that the usage of the label assigned to the test construct should be determined by its target domain. In the context of the CET-4, however, it is determined by policy. The inconsistency between the narrowly defined target domain and the broadly interpreted label of English proficiency is associated with misuses and abuses of the test beyond the educational context. I will take up this point in Section 6.4.

The extrapolation inference extends the interpretation of the CET-4 score from the universe of generalization to the target domain of the test. Kane (2006) maintains that the extrapolation inference in IUA can be evaluated in terms of the relationship between the universe of generalization and the target domain. Findings of this study provide analytic and empirical evidence for the relationship between students’ CET-4 scores and estimates of their performance over the target domain defined by the College English curriculum. Deyi’s account of the connection and overlap between the CET-4 and the
CECR, as well as College English teaching and learning (see Section 4.2.2) provide analytic evidence for the extrapolation inference. The account by Edi of the consistency between students’ CET-4 score and their performance in the College English classroom (see Section 4.4.2) provides empirical evidence for this inference. These results indicate that the alignment between the test design and the College English curriculum, coupled with the quality control procedures discussed above, is perceived to make the CET-4 score a valid measure of students’ achievement of the curriculum.

Interpretability of test score is a critical issue related to the utilization inference in the IUA. Test users may choose to rely on information provided by the test developer in interpreting the test score. Therefore, the test developer always has the responsibility to provide support for the proposed score interpretation (Kane, 2013; Shepard, 1997). In order to enhance the interpretability of the CET-4 score for test users, the NCETC provides detailed explanations of the reported norm-referenced test scores as well as information regarding the normalization process. However, there is inconsistency between the CET-4 as a criterion-related English proficiency test and the way the test scores are reported. Language proficiency test scores nowadays tend to be interpreted with reference to criteria regarding language use abilities in context (McNamara, 2010). This criterion-referenced approach to assessment contrasts with the norm-referenced approach used in the CET-4 context, which interprets an individual’s score with reference to the performances of others. The goal of norm-referenced tests is to rank test takers rather than to determine whether or not they have demonstrated mastery of a certain skill or set of skills. Therefore, even if test users could interpret the reported percentile CET-4 score in the way intended by the test developer, the score would still be not meaningful to
them in knowing whether or not, or to what extent, the students meet the requirements specified in the curriculum. In this sense, the usefulness of the CET-4 in achieving the intended consequence of informing teaching and learning would be jeopardized.

The findings of this study highlight the CET-4 developer’s efforts to achieve the intended consequences of promoting College English education by providing useful information about students’ English proficiency based on the curriculum. However, the large test taker population of the CET-4 and the other forces operating on the Chinese educational scene impose challenges for the achievement of the intended consequences and for minimizing unintended consequences of the test by improved test design. These challenges are best explained by Deyi in her interview (see Section 4.2.2):

But we can only make limited changes in terms of test design; we cannot make dramatic changes. Otherwise, we would be overwhelmed by complaints and criticism. You know, for large-scale testing, there are a great many constraints and we have to strive to survive and compromise with various competing forces.

Given these challenges, the CET developer gives priority to the scoring and generalization inferences in the IUA, particularly through the quality control procedures in marking and the standardization process. The purpose is to ensure that all test takers are asked to perform the same tasks under the same conditions so that no test taker would be potentially disadvantaged by factors such as test content, test condition, response mode and format, as well as scoring of constructed-response items. This purpose is relevant to the conception of *procedural fairness* (Kane, 2010). Procedural fairness is concerned with how test takers are treated, in particular with how consistently and fairly they are treated in the testing process, and is, therefore, largely under control of the test
developer. This notion of fairness is also related to what McNamara and Ryan (2011, p.163) called test fairness for individual candidates: “the extent to which the test quality, especially its psychometric quality, ensures procedural equality for individual and subgroups of test takers.” McNamara and Ryan (2011) pointed out that the concern for fairness in this sense corresponds to the evidential basis of test score interpretation and test use in Messick’s (1989) validity matrix. This very basic notion of fairness leaves out many issues related to consequences and value implications of test use and interpretation.

The value of fairness in the sense of equality is rooted in the long history of large-scale high-stakes testing in China as introduced in Chapter 1. Due to this history, fairness via examinations is the foundation upon which testing and examinations function in Chinese society (Cheng, 2009). Findings of this study indicate that reliability is given priority compared with other qualities of the CET-4, and procedural fairness is the core value held particularly important by the CET-4 developer, given the large scale of the test as well as the increase in the number of constructed response questions in the test.

In summary, despite the test developer’s efforts, there is evidence from both this study and other studies (e.g., Chen, 2011; Wang & Sun, 2012) that constitutes rebuttals against the decision rule and the utilization inference in the CET-4 IUA. For example, standardization may be a source of construct-irrelevant variance, and the inconsistency between the target domain of the test and the construct description constitutes construct underrepresentation. The reported norm-referenced test score may jeopardize the usefulness of the test in informing teaching and learning about students’ achievement against the curriculum. Moreover, the CET-4 is administered semi-annually at the end of each semester. It summarizes learning related to general criteria (i.e., statement of
attainment in the national curriculum) and does not have the detail that enables it to be
diagnostic to the degree needed to help specific teaching and learning. The CET-4 as an
externally and centrally administered test exerts considerable pressure on its stakeholders
such as teachers and students teaching and learning English in Chinese universities and
colleges. These limitations present tremendous challenges for the CET-4 developer to fit
the test to its purpose of informing teaching and learning.

6.3 Uses of the CET-4 in the Educational and Societal Contexts

The results reported in Section 4.3 and Section 4.4 show multiple uses of the
CET-4 in educational, business, and government contexts, as well as the test users’
justifications for their test use decisions. These findings, together with findings of other
studies (e.g., Shu, 2004; Wang & Sun, 2012; Zhang, 2005) reveal the power and impact
of the CET-4 in Chinese educational and societal contexts.

6.3.1 Uses of the CET-4 in the Educational Context

The uses of the CET-4 in the educational context include both direct and indirect
uses. Direct use involves making high-stakes decisions about students based on their test
scores. Indirect use capitalizes on the motivational effects of testing or uses testing and
test reporting to shape public opinion (Haertel, 2013). Results reported in Section 4.3.1
reveal two major indirect uses of the CET-4 in the College English education program.
These uses are consistent with the intended purposes of the test.

First, the CET-4 results were and continue to be used to inform College English
teaching in both Edi’s and Edu’s universities. The reported sub scores were found
particularly useful in terms of providing pertinent information for College English
teaching. In this sense, the CET developer’s effort to enhance the usefulness of the test by
reforming the score reporting system achieved its purpose. However, as mentioned above, the CET-4 is a norm-referenced test, and norm-referenced scores are useful for ranking students. Using the norm-referenced CET-4 results to inform teaching and learning requires further validation inquiries. Moreover, the CET-4 is a summative test administered to students usually after studying for two years in their College English education programs, during which they have probably been taught by different teachers. Results from the test provide information about students’ learning at the aggregate level. Using these results as sources of information about specific classrooms to guide instruction poses concerns for validity as well.

Second, there is the indirect consequence of the CET-4 in terms of enhancing the College English education program. For example, in Edi’s university, financial support to the program was increased and the status of the program was promoted as a consequence of using the CET-4 to evaluate quality of universities across the country. This use is relevant to the intended consequences of the CET-4 implementation to promote tertiary EFL education and to improve higher education in China. However, evaluation of university quality is not the use for which the CET-4 is designed. This use is more related to the educational context in which the test is implemented and not traceable to test design.

The CET-4 score was used in both Edi’s and Edu’s universities as a prerequisite for the undergraduate students to obtain their bachelor’s degrees. This use has high stakes for students and is unintended by the CET-4 developer. In order to mitigate the high stakes of the CET-4 in the educational context, the CMoE clearly indicates that universities are not to link graduation or degrees to the CET-4 results and should use the
results together with information from other sources to evaluate students’ English proficiency. In these two universities, internal English tests were administered. However, the purpose of these tests was to alleviate the high stakes attached to the CET-4 score, rather than to collect information about students’ English proficiency. The rationale of using the CET-4 for degree conferment is to motivate students to put more effort into studying English. Motivation is often considered an intended consequence of large-scale testing (e.g., Anderson et al, 1990; Burger & Kroeger, 2003; Deci & Ryan, 2002; Han & Yang, 2001; Wang, 2008; Yeh, 2005). Findings of this study highlight the test users’ dilemma in using the CET-4 in the educational context for the purpose of motivation. On the one hand, the motivational function of the test is associated with its high stakes. On the other hand, the test users are concerned about the impact of the test on students’ well being.

6.3.2 Uses of the CET-4 in the Business and the Government Contexts

In the government context, the CET-4 score was used in City A to control access to social resources and in City B as a gatekeeper to government positions. In the business context, the score was used as a tool to screen candidates in the recruitment process in Buna’s and Buni’s companies although English was not relevant to the workplace. Compared with the uses in the educational context, these uses have even higher stakes for students. Test users in these contexts sought to justify their test use decisions based on their extended interpretations of the CET-4 score: an indicator of students’ English proficiency, which, in turn, is regarded as a proxy of student quality. In Buran’s and Buru’s companies, where English was relevant to the workplace, the CET-4 score was used in the recruitment process together with evidence of candidates’ English proficiency
from other sources. However, the rationale underlying this test use is not the relevance of the CET-4 score to the workplace. To Buran and Buru, the CET-4 score is a limited representation of students’ English use abilities. They used the terms such as academic/school/textbook English to refer to the testable knowledge and skills of English measured by the CET-4. Their domain of interest is English language use in the workplace. As the other test users in this study, Buran and Buru sought to justify their test use decisions based on the extended interpretation that the CET-4 score indicates if the student reached the basic university requirement.

The actual uses of the CET-4 in the above contexts lent support to Fulcher and Davidson’s (2007) argument that language tests used as policy instruments are usually co-opted into services for which they were not designed. The CET-4 is designed to measure students’ English proficiency based on the basic requirements specified in the College English curriculum. However, its results are used as a gatekeeper to access to resources in various societal domains such as workplace, government, and social insurance. Associated with these direct uses of the test are the high stakes for students. A single assessment program used to serve multiple purposes poses concerns for validity (Klinger et al., 2008; Koch & DeLuca, 2012; Shulha & Wilson, 2009).

Decisions based on inaccurate interpretations are unacceptable ethically, legally, and socially and tend to bring about adverse consequences (Kane, 2013; Messick, 1996). To address this issue, the test developer and test users should take shared responsibilities (Xi, 2010). On the one hand, the test developer needs to provide accurate information to test users. On the other hand, test users are accountable for the consequences of the decisions they make. Indeed, the CET-4 results might have been used by the test users to
justify decisions they had already made. The findings discussed above echo the impact of the CET-4 in Chinese society, which has already been raised by other researchers (e.g., Zheng & Cheng, 2008). To better understand why the CET-4 remains such a powerful test in China, it is necessary to consider the values embedded in the test construct and held by its test users.

6.3.3 Value Implications Underlying Actual Uses of the CET-4

The justifications test users in this study provided for their test use decisions reveal two major types of values: 1) values embedded in the test construct, and 2) values associated with the test per se.

The extended interpretations of the CET-4 score, based on which the test users seek to justify their test uses, highlight the social values embedded in the test construct. In the educational context, the CET-4 score is used for degree conferment with the purpose of motivating students to study English. This use is based on the extended interpretation that the test score represents students’ efforts rather than their English proficiency. This is not surprising given that achievement is believed to depend upon effort rather than ability alone and that examinations, particularly high-stakes examinations, are often used within the Chinese context to motivate students to study (Han & Yang, 2001; Wang, 2008).

In the societal context, high-stakes test use decisions are made based on the implications associated with the construct of English proficiency rather than the relevance of the test score to the test use context. The CET-4 score is associated with qualities of students. This association is consistent with the CMoE’s emphasis on the quality of China’s higher education in general and quality of College English education
The CET-4 testing program is implemented to promote the quality of EFL education in the Chinese tertiary education context in the service of political, social, and educational needs. In this context, the test construct is authorized by policy and has social value implications embedded in the Chinese educational and sociopolitical contexts. The findings discussed above lend support to Koch and DeLuca’s (2012) argument that large-scale assessment operates within and also has an influence on the specific context of practice where the various interpretations and uses of its data are situated. Findings of this study point to the importance of contextual factors in validation. As Moss et al. (2006, p. 145) argued, “externally mandated tests are always interpreted and used in particular local context, which shape and are shaped by them.” In language assessment, McNamara and Ryan (2011) assert that value implications in test construct and its use are more of a justice issue, which concerns the consequential basis for test score interpretation and test use in Messick’s (1989) terms. They maintain that issues of social value should be investigated in light of cultural analysis and critical policy analysis, considering questions of history, ideology, and discourse context. Therefore, investigations of the consequences of the CET-4 need to go beyond the micro-level of test construct into the macro-social context by investigating and articulating values and power dimensions implicit in the test. This thesis study is one of the few attempts to do so empirically.

Findings concerning actual direct uses of the CET-4 score also highlight values associated with the test per se. The test users seek to justify their high-stakes decisions based on their perceived credibility, authority, and quality of the CET-4. These perceptions are related to the large scale, the status of the developer, the long history of
the CET-4 test, as well as the traditional acceptance of large-scale testing in China. For example, in the societal context, the CET-4 is regarded as a test with high credibility, authority, and quality because it is a national large-scale standardized test administered by the CMoE and is required of all the university undergraduates across the country. The value of the CET-4 for these users might be that it could make the recruitment appear to be an objective and fair process. In this context, the test had unchallenged authority. The test developer’s efforts to enhance the quality of the CET-4 were effectively politicized, and the technical quality of the test became a means of enforcing power. Tests can disarm criticism (McNamara, 2010; McNamara & Roever, 2006; Shohamy, 2009).

From the validity perspective, however, the perceived credibility, authority, and quality of the CET-4 do not justify the extrapolation inference in the IUA, which is based on the assumption that the test score provides information relevant to the test use domain. The quality control procedures by the CET-4 developer such as standardization are politicized. In this sense, the technical quality of the test becomes a means of enforcing power. This is particularly true in the Chinese context, where large-scale high-stakes examinations have traditionally been used as a means for selection purposes. As elaborated in Section 1.3, examinations such as Keju and Gaokao and the CET-4, despite their differences in design and intended uses, have all been recognized and accepted in Chinese society as a justifiable means to control access to limited resources in various societal domains such as government, education, and workplace. Cheng (2009) asserts that examinations continue to enjoy a societal acceptance and recognition in China as a fair measurement for selection. Findings of this study are consistent with this assertion.
Overall, this study unpacked the impact of the CET-4 in the educational and societal contexts and fleshed out the values underlying the multiple interpretations and multiple uses of the test. Findings of this study challenged the validity of the test and bring ethical issues to the foreground. Clearly, the defensibility of the values underlying the CET-4, both in its construct and in its use, is an issue that is open to dispute.

6.4 Washback of the CET-4 on Learning

Findings about washback on students’ learning provide particularly important evidence regarding consequences of the CET-4, given its purpose of promoting College English education. This section discusses the washback effects of the test on students’ learning, in relation to the test developer’s efforts supporting the IUA discussed in Section 6.2 and the actual uses of the CET-4 discussed in Section 6.3. This discussion explains the possible sources of the washback from the SRL and the sociocultural perspectives of learning.

6.4.1 Intended and Unintended Washback of the CET-4 on Learning

Results reported in Chapter 5 show the following CET-4 washback on students’ learning that is consistent with the intended consequences of the test by the test developer.

First, students’ perceptions of the CET-4 demands are a significant variable that influences their learning practices (standardized path coefficient=.52, p<.05). The higher the students’ perceived CET-4 demands are, the more likely they will engage in test preparation. These results are compatible with results from previous studies showing that students’ perception of test demands can lead to learning action changes (Green, 2007; Shih, 2007; Xie, 2010). These findings suggest that a test can direct students’ efforts on
learning the knowledge and skills that it measures, implying the possibility of promoting learning by means of improved test design (Messick, 1996)

Second, as shown in Figure 5-2, students’ perceptions of the CET-4 demands are fundamentally represented by four indicators: perception of English language knowledge and skills (LKS), perception of listening skills (LIS), perception of background knowledge (BAK), and perception of reading skills (RES). Statistical results presented in Figure 5.2 and Table 5.19 show that perception of LKS is the strongest indicator, followed by perception of RES and perception of LIS. Descriptive statistics reported in Table 5.3 indicate that in general, students in this study perceive that the CET-4 has high demands of LKS (with a mean of 6.07 out of 7). The perceptions of the demands for the reading and listening skills are also high (5.82 and 5.56 out of 7 respectively). These findings are consistent with the construct the CET-4 is intended to measure (i.e., students’ English language knowledge and skills, and reading comprehension and listening comprehension are two major sections of the test). In this sense, findings of this study lend support to the construct validity of the CET-4 and provide evidence for the intended consequence that the CET-4 score provides relevant information about students’ English proficiency based on the curriculum.

Third, results summarized in Table 5.6 indicate that students’ perceptions of the achievement use of the CET-4 are high (5.47 out of 7). This is consistent with the test developer’s intention to promote English language academic achievement of the students. However, this result is inconsistent with Xie’s (2010) study on the CET-4, which found that out of the 895 student participants, only 25.7% students took the CET-4 for the purpose of achievement-oriented test uses. To understand and explain this discrepancy,
further investigations need to be conducted to compare the contexts where the data were collected for the studies on the same test.

Fourth, results reported in Table 5.12 and Figure 5-2 indicate that students in this study engaged in four major types of test preparation practice for the CET-4: long-term skill development, rehearsal and cramming, test preparation management, and social-affective strategies. Of these four strategies, the social-affective strategy was used most often by the students (5.45 out of 7). This result concurs with findings from studies on students’ CET-4 preparation strategies from the SRL perspective (e.g., Xiao, 2014). The CET-4 seems to have had a significant motivational effect on students’ learning in terms of the use of this strategy.

Results reported in Chapter 5 also reveal unintended consequences of the CET-4 on students’ learning. First, students’ perceptions of instrumental uses of the test are high (5.26 out of 7). This is not surprising given the high stakes of the CET-4 in both the educational and societal contexts, as indicated by results reported in Chapter 4. Findings of this study indicate that despite the CET-4 developer’s efforts to mitigate the high stakes attached to the test, its results are still being extensively used in both the educational and societal contexts for high-stakes decisions. Associated with these stakes is the pressure on students. Harlen (2006) asserts that the pressure exerted by the external testing requirements is a factor that decreases the usefulness of information provided by summative assessment for improving teaching and learning. In this sense, the effectiveness of using the CET-4 to improve teaching and learning is limited.

Second, the SEM results indicate that compared with the influence of students’ perceptions of the CET-4 demands, their perceptions of the test uses have a more
powerful effect on their test preparation practices (standardized path coefficient=.92, 
$p<.001$). Statistics in Table 5.17 also show that the correlation of perceptions of 
instrumental use with the use of the rehearsal and cramming strategy ($r=.57$) is higher 
than its correlation with the use of long-term skill development strategy ($r=.39$). An 
example of the use of the rehearsal and cramming strategy is that students memorize 
model essays to prepare for the CET-4 writing, as mentioned by both Deyi and Edu in the 
interviews. These results are consistent with findings of other studies (e.g., Xiao, 2014; 
Xie, 2010) that students who took the test for instrumental goals tend to use cue-seeking 
and narrow-focusing test management strategies. Related to the instrumental use of the 
CET-4 score are the high stakes for students such as degree conferment and employment. 
Raising the stakes for students increases their extrinsic motivation (Green, 2007). 
“Extrinsically motivated learners were more likely to take shortcuts to the extrinsic goal, 
sacrificing learning in the process by using strategies that ensured short-term but not 
goals are associated with the utilitarian values of learning another language. Chen (2011) 
attributed students’ perception of the instrumental uses of the CET-4 to the Chinese 
culture of learning, which is characterized by examination orientation and a utilitarian 
function. She argued that this learning culture may militate against the successful 
implementation of the current College English Reform Program in China. Therefore, 
irrespective of the quality of the CET-4 in terms of construct validity, the effectiveness of 
using the test to enhance students’ learning is limited.

Third, as shown in Table 5.19, the path coefficient estimate for the 
$Prep\rightarrow Perform$ path is very small and insignificant at the .05 $\alpha$-level. These results
indicate that students’ test preparation is not a significant predictor of their test performance. Other washback studies on the CET-4 (e.g., Xie, 2010) and other tests (e.g., Green, 2007; Perrone, 2010, 2011) also found that test preparation does not contribute much to improving performance. Therefore, there are limitations of using standardized tests such as the CET-4 to improve students’ achievement.

6.4.2 Factors Contributing to the Washback of the CET-4

Findings of this study indicate that the complexity of the CET-4 washback is that students’ test preparation practices are influenced by both test properties and contextual factors. Test preparation is an active, constructive SRL process, “whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment” (Pintrich, 2000, p. 453). In this process, motivational characteristics interact with strategic behaviors and should be best understood from the sociocultural perspective (Rose, 2012; Zimmerman & Kitsantas, 2005).

Students’ perceptions of test demands were found to be a significant influencing factor on learning. From the SRL perspective, perceptions of test demands may prompt students to strategically use their resources and to focus their efforts on the target skills necessary to the test tasks during test preparation. This perception is similar to the concept of “task knowledge” regarding “how to go about doing a particular task and the knowledge and skills needed to do so” (Wenden, 1998, p. 518). Task knowledge is one type of metacognitive knowledge, which provides the knowledge base for effective planning, monitoring, and evaluating in the learning process (Wenden, 1998). From this perspective, overlap between the CET-4 and the curriculum is important in inducing
intended washback in terms of focusing students’ efforts on the knowledge and skills measured by the test. The more the test design overlaps with the curriculum, the more likely it is that positive washback would be generated (Green, 2007).

However, results reported in Chapter 5 show that a long-term skill development strategy—which is “believed to be beneficial for language learning in the long-run and to help maintain healthy interest towards English learning” (Xie, 2010, p. 234)—is least used by the students. These results are consistent with findings of Xiao’s (2014) CET-4 washback study and are related to the limitations of the CET-4 as a large-scale standardized test in assessing learning. The CET-4 includes a limited range of tasks in restricted language use contexts. Students may focus their efforts on test-taking skills necessary to answer the CET-4 questions at the expense of long-term English language skill development. Crooks (1988) pointed out that skills that are easier to test may come to be better represented on tests than skills that are perhaps equally important but not easily tested. Teachers and students may be able to predict the CET-4 test content and work efficiently to improve the test score by directing their efforts to the tested areas of the curriculum at the expense of untested areas. This issue has also been discussed in other CET-4 washback studies (e.g., Gu, 2005; Ren, 2011). Therefore, due to the restrictions in the CET-4 design, the function of the test in terms of improving students’ English proficiency is limited.

The sociocultural perspective construes learning in the interaction between learners and their environments (Moss et al., 2006). Discussion in Section 6.3 highlights the powerful role the CET-4 is playing in the learning environment. The motivational function of the CET-4 relies, to a great extent, on the high-stakes uses of the test in both
the educational and societal contexts. These uses are based on the test users’ extended interpretations of the test score and give rise to ethical issues because of the serious unintended consequences. Therefore, to better understand why the CET-4 washback exists in the Chinese educational context, it is necessary to analyze not only the micro classroom context, but also the macro Chinese educational and sociopolitical systems.

6.5 Conclusions

The CET-4 is intended to achieve two interrelated goals: to provide useful information about students’ English language proficiency based on the curriculum; and 2) to promote EFL education in the Chinese tertiary educational context. The test achieved its intended consequences in terms of providing test users in the educational context with accurate and relevant information about students’ achievement of the College English curriculum. The accuracy of the information is related to the alignment between the test and the curriculum as well as the CET developer’s efforts to support the scoring and generalization inferences in the IUA. Recent reform of the score reporting system contributes to the relevance of the CET-4 scores to College English education.

With regard to the second goal, findings of this study provide evidence that some intended washback effects of the CET-4 are being achieved. At the program level, the test scores are used by educators to inform teaching. Also, financial input into the College English programs is increased, which contributes to the promotion of College English education. For students, the CET-4 plays a significant motivational function in terms of directing their efforts on learning the knowledge and skills measured by the test. Students’ perceptions of the CET-4 demands are a significant factor that influences their
learning processes. These findings suggest the possibility of enhancing learning through improving test design.

However, the SEM results indicate that test preparation is not a significant factor influencing students’ learning outcomes. Therefore, although the CET-4 program plays a fundamental motivational role in terms of focusing students’ efforts on the tested areas, its function in terms of improving students’ achievements may be limited. Moreover, findings of this study highlight the following two unintended consequences of the test.

First, the CET-4 is used by various users in different contexts for multiple purposes such as degree conferment, employment, and allocations of social resources. These uses go beyond the intended uses the test is designed for and are embedded in the Chinese educational, social, and political contexts. They raise stakes attached to the CET-4. High-stakes test use decisions based on extended interpretations of the test score create critical ethical and validity issues.

Second, with regard to washback of the CET-4, students’ perceptions of test uses have a powerful influence on their test preparation. This perception constitutes extrinsic motivation that directs students’ efforts on the tested areas of the College English curriculum. Due to practicality, there are limitations of the CET-4 test design in terms of the limited tasks and the limited range of language use context included in the test. These limitations may induce negative washback on students’ learning (i.e., focusing only on the tested areas of the curriculum and test-taking skills at the expense of long-term English language proficiency development). Therefore, the function of the CET-4 to improve students’ English proficiency is limited.
In conclusion, this study provides evidence for the construct validity of the intended CET-4 score interpretation as a general measure of students’ English proficiency based on the College English curriculum. The CET-4 score provides accurate, relevant, and useful information about students’ achievement of the curriculum. This evidence is related to the CET-4 developer’s efforts to enhance the quality of the test. However, this evidence does not, in itself, justify uses of the test, which have to be evaluated in terms of intended and unintended consequences. Due to the limitations of the CET-4 in terms of test design, the intended consequence of promoting learning is achieved to a limited extent. Washback of the CET-4 on students’ learning is traceable to the test design, but more related to the Chinese educational and sociocultural contexts in which the test is implemented and functioning. The motivational function of the CET-4 relies more on the high-stakes uses of the test. These uses imply serious unintended negative consequences for students, and give rise to validity as well as ethical concerns because justification of the test use decisions was based on the circular argument rather than the relevance of the test construct to the test use context. Therefore, the decision rule in the IUA for the CET is not supported by findings of this study.

6.6 Limitations and Directions for Future Research

One major limitation of this study is its limited scope due to logistical and time restrictions. Eight test user participants from four different contexts were recruited in two cities in this study and only one interview was conducted with each of them. It would be ideal to do more follow-up interviews. Student participants were sampled from second-year undergraduates in two universities. Given the restrictions in participants and
contexts, generalization of the findings of this study to other contexts should be undertaken with caution.

A second limitation of the study concerns the collection of self-reported data using a questionnaire. It is problematic for a questionnaire to capture the complexity of the CET-4 washback on students’ learning.

Another limitation is the lack of real students’ test performance data. Despite my effort to obtain genuine CET-4 data, I was not granted access to the actual data by the CET developer. Students’ CET-4 performance was measured in this study with a researcher-administrated CET-4 through combinations of previous CET-4 tests, although students indicated that they had not had a chance to review these test papers. Moreover, due to time restrictions, post-preparation test data were not collected to compare with the pre-preparation test data so that the effects of test preparation could be examined more accurately.

For future research, first, studies on the consequences of the CET-4 should involve more test stakeholders (e.g., educational administrators, parents, and test users in more diverse contexts) to provide further in-depth descriptions of the multiple interpretations and multiple uses of the test as well as the interactions among them.

Second, considering the social dimension of language testing highlighted by this study, future research on the impact and washback of large-scale high-stakes language tests should be conducted based on systematic analysis of the broader social context, drawing on insights from disciplines such as sociology and policy analysis.

Finally, given that consequences are a function of numerous factors in the context where the test is implemented and functioning, in the test, and in the participants, future
research requires collaboration of researchers from different fields such as educational measurement, applied linguistics, psychology, sociology, and policy studies.

### 6.7 Contributions and Implications for Test Development

This study has made original contributions to language assessment research.

First, drawing on the argument-based approach to validation, the study synthesized findings concerning the intended and unintended consequences of the CET-4 and other evidence about test validity, particularly the test developer’s efforts to support different inferences in the IUA. It is a timely attempt to investigate impact and washback within a coherent validation framework.

Second, this study investigated washback in both a validation framework and the broader social context. Compared with previous washback studies, it provides a more comprehensive perspective on understanding this complex phenomenon: what it looks like, how it works, and why it exists in the particular context. Methodologies used in this study and findings of this study have significant implications for future washback and impact research both in the CET context and other testing contexts.

Third, by investigating the value implications underlying the CET-4 construct and the test implementation, this study highlights the social dimension of language testing. Evidence about test consequences in relation to the social context provides insights for policy implementation.

Fourth, this study highlights the importance of systematically analyzing the contextual factors that contribute to test consequences. EFL teaching and learning are self-regulated processes situated in the educational and sociocultural contexts. Systematic analysis of these contexts will provide valuable information for the test developer to
achieve the intended consequences of improving College English education by implementing the CET-4.

Results regarding the CET-4 developer’s efforts to achieve the intended consequences of the test highlight the challenges of enhancing the technical qualities of the test while meeting the educational and social needs within the practicality of the large-scale test. Findings of this study suggest some limitations of the CET-4.

First, the test is described as a criterion-related, norm-referenced test. This description is confusing and the reported norm-referenced scores do not provide much useful information for test users in the educational context about students’ achievements of the curriculum. Given the purposes and design of the test, the CET-4 should be described as a criterion-referenced test based on the College English curriculum.

Second, the CET-4 is designed to measure students’ English language proficiency, and this construct is given a broad and expansive interpretation in the College English curriculum. Given the restriction of the CET-4 in test design, there is the issue of construct underrepresentation. To address this issue, the CET-4 should be re-defined as an achievement test based on the curriculum.

Third, although many of the unintended uses of the CET-4 are beyond the control of the test developer, the CET developer should make more specific and clearer claims about how the test can be used and about the benefits for such uses. Given the impact of the CET-4 in society and education, the CET-developer should conduct investigations on the consequences associated with some common uses of the test. As Nichols and William (2009) suggest, a test developer’s responsibilities to collect test score use evidence should expand to encompass consequences associated with test score uses that become
increasingly common over time. This responsibility is particularly relevant in the context of the CET-4 where the test is being used as policy instrument for educational reform.

In conclusion, the CET-4 is a testing program initiated and implemented to serve the needs of the Chinese educational system and of Chinese society at large. The implementation of the CET-4 is a contributing factor to the context in which the test results are interpreted and used. As a consequence of this policy implementation, the status of the English language is promoted. Therefore, the CET-4 is playing a significant mediating function in enforcing de facto EFL policy in China. Test users capitalize on this status as well as the test developer’s efforts for the technical qualities of the test to justify their test use decisions. This study fleshes out the washback and impact of the CET-4 in the Chinese educational and societal contexts. It highlights the power and values implied in the test construct and underlying its use and implementation, thus, reveals the social dimension of the test.

The need to understand and evaluate consequences of the testing program in a justifiable way presents significant challenges for the policy-maker, the tester, and the researcher. In order to maximize beneficial effects of the program and to minimize its detrimental effects, collaborations of these different stakeholders are required in future research.
References


Appendix A. Ethics Clearance Approval Letter

September 25, 2011

Mr. Youyi Sun
Ph.D. Candidate Faculty of Education
Duncan McArthur Hall, A106
Queen’s University
511 Union Street
Kingston, ON K7M 5R7

GREB Ref #: GEDUC-578-11; Romeo # 6006274
Title: "GEDUC-578-11 Context, Construct, and Consequences: Washback of the College English Test in China"

Dear Mr. Sun:

The General Research Ethics Board (GREB), by means of a delegated board review, has cleared your proposal entitled "GEDUC-578-11 Context, Construct, and Consequences: Washback of the College English Test in China" 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,

Joan Stevenson,
Ph.D. Professor and Chair
General Research Ethics Board
Appendix B. Consent Form to Test Developer Participant

CONTEXT, CONSTRUCT, AND CONSEQUENCES:
WASHBACK OF THE COLLEGE ENGLISH TEST IN CHINA

Name (please print clearly): ______________________________________

1. I have read and retained a copy of the Letter of Information and Consent Form and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Context, Construct, and Consequences: Washback of the College English Test in China. I understand that the purpose of this study is to investigate the College English Test (CET) developer’s score interpretation and the intended score use in comparison with and in relation to various test users’ interpretations of the test scores and their uses of the test scores in both the educational and societal contexts. I understand that participating in this study involves checking the accuracy of the interpretive argument for the CET constructed by the researcher and making comments on it, which takes about five hours.

3. I understand that my participation in this study is voluntary and I may withdraw at any time. If I withdraw, I may request removal of all or part of my data. I understand that every effort will be made to maintain the confidentiality of the data now and in the future. Only the researcher and his supervisor will have access to the data. The data may also be published in professional journals or presented at academic conferences, but any such presentations will be of general findings and will never breach individual confidentiality. Should I be interested, I am entitled to a copy of the findings.

4. I am aware that if I have any questions about study participation, I may contact the researcher Youyi Sun at youyi.sun@queensu.ca; his supervisor, Dr. Liying Cheng at (01) 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 (01) 613-533-6081 or chair.GREB@queensu.ca.

I have read the above statements and freely consent to participate in this research.

Please sign one copy of this Consent Form and return to your classroom teacher. Retain the second copy for your records. If you wish to receive a copy of this study, please provide your e-mail address below.

Signature: ___________________________ Date: ______________________

Email address: ______________________
Appendix C. Consent Form to Test User Participants

CONTEXT, CONSTRUCT, AND CONSEQUENCES:
WASHBACK OF THE COLLEGE ENGLISH TEST IN CHINA

Name (please print clearly): _________________________________

1. I have read and retained a copy of the Letter of Information and Consent Form and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Context, Construct, and Consequences: Washback of the College English Test in China. I understand that the purpose of this study is to investigate the College English Test (CET) developer’s score interpretation and the intended score use in comparison with and in relation to various test users’ interpretations of the test scores and their uses of the test scores in both the educational and societal contexts. I understand that participating in this study involves participating in a semi-structured interview at a place of my choice which takes about 45 minutes.

3. I understand that my participation in this study is voluntary and I may withdraw at any time. If I withdraw, I may request removal of all or part of my data. I understand that every effort will be made to maintain the confidentiality of the data now and in the future. Only the researcher and his supervisor will have access to the data. The data may also be published in professional journals or presented at academic conferences, but any such presentations will be of general findings and will never breach individual confidentiality. Should I be interested, I am entitled to a copy of the findings.

4. I am aware that if I have any questions about study participation, I may contact the researcher Youyi Sun at youyi.sun@queensu.ca; his supervisor, Dr. Liying Cheng at (01) 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 (01) 613-533-6081 or chair.GREB@queensu.ca.

I have read the above statements and freely consent to participate in this research.

Please sign one copy of this Consent Form and return to the researcher. Retain the second copy for your records. If you wish to receive a copy of this study, please provide your e-mail address below.

Signature: ________________________________ Date: __________________

Email address: ______________________________
Appendix D. Consent Form to Student Participants

CONTEXT, CONSTRUCT, AND CONSEQUENCES:
WASHBACK OF THE COLLEGE ENGLISH TEST IN CHINA

Name (please print clearly): _______________________________________

1. I have read and retained a copy of the Letter of Information and Consent Form and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Context, Construct, and Consequences: Washback of the College English Test in China. I understand that the purpose of this study is to investigate the influence of the College English Test (CET) in China on students and their learning within the Chinese tertiary context in relation to test design and test use. I understand that participating in this study involves completing a questionnaire in the classroom which takes about thirty minutes and taking a College English Test (CET) Band 4 test in the classroom which takes about 125 minutes.

3. I understand that my participation in this study is voluntary and I may withdraw at any time with no effect on my standing in school. If I withdraw, I may request removal of all or part of my data. I understand that every effort will be made to maintain the confidentiality of the data now and in the future. Only the researcher and his supervisor will have access to the data. The data may also be published in professional journals or presented at academic conferences, but any such presentations will be of general findings and will never breach individual confidentiality. Should I be interested, I am entitled to a copy of the findings.

4. I am aware that if I have any questions about study participation, I may contact the researcher Youyi Sun at youyi.sun@queensu.ca; his supervisor, Dr. Liying Cheng at (01) 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 (01) 613-533-6081 or chair.GREB@queensu.ca.

I have read the above statements and freely consent to participate in this research.

Please sign one copy of this Consent Form and return to your classroom teacher. Retain the second copy for your records. If you wish to receive a copy of this study, please provide your e-mail address below.

Signature: ________________________________  Date: __________________________

Email address: __________________________
Appendix E. Interview Guideline with the CET Developer Participant

1. The CET Syllabus describes the CET test score as criterion-related and norm-referenced. Could you please explain this in more detail?

2. How are the CET-4 papers marked?

3. How can you ensure the reliability and validity of the CET-4?

4. Why is reliability/validity/fairness so important for the CET-4?

5. What is the intended interpretation of the CET-4 score?

6. What are the intended uses of the CET-4 results?

7. What do you think of the consequences of the CET-4?

8. How could the CET-4 provide more useful information to its intended test users?
Appendix F. Interview Guideline with Test User Participants

A. Personal and business information
1. How many employees are there in your company?
2. What does your company do?
3. How many undergraduate students apply to work in your company each year?
4. How many new employees does your company recruit each year?
5. What is your position?
6. How long have you been working in this position?
7. Have you ever taken the CET?

B. CET score interpretations and uses
1. Do you agree that the reliability of the CET is high because of its quality control in terms of scoring?
2. Are there any sectors or divisions in your company that require employees with a certain level of English proficiency?
3. Do you require CET scores from undergraduates applying to work in your company? Why (not)?
4. Do you think an undergraduate with a higher CET score would have more opportunities to work in your company?
5. Do you think an undergraduate with a high CET score would use English proficiently in the workplace?
6. What are you looking for when you require the CET score from the job applicant?
7. What is your interpretation of the CET score?
Appendix G. First Page of CET-4 sample Test Paper

College English Test Band 4

Part I  Writing (30 minute)

Directions: For this part, you are allowed 30 minutes to write a campaign speech in support of your election to the post of chairman of the student union. You should write at least 120 words following the outline given below in Chinese:

1. 你认为自己具备了什么条件（能力、性格、爱好等）可以胜任学生会主席的工作?

2. 如果当选，你将为本校学生做些什么?

A Campaign Speech

Part II Reading Comprehension (Skimming and Scanning) (15 minute)

Directions: In this part, you will have 15 minutes to go over the passage quickly and answer the questions on Answer sheet 1.

For questions 1-7, mark

Y (for YES) if the statement agrees with the information given in the passage;

N (for NO) if the statement contradicts the information given in the passage;

NG (for NOT GIVEN) if the information is not given in the passage

For questions 8-10, complete the sentences with the information given in the passage.
Appendix H. Questionnaire for Student Participants

The purpose of this questionnaire is to investigate your perceptions of the CET-4 and your test preparation process. There are no right or wrong answers; you only need to choose ONE option that is the closest to your situation. There are five sections in the questionnaire.

Section A. Your perceptions of the CET demands. This section includes statements about the skills and/or knowledge necessary for answering different CET tasks. Please circle ONE number to indicate the extent to which you agree with each of the statements.

- 1 = Strongly disagree • 2 = Disagree • 3 = Somewhat disagree
- 4 = Neither agree or disagree • 5 = Somewhat agree • 6 = Agree • 7 = Strongly agree

**In order to do well in the Listening Comprehension section,**

<table>
<thead>
<tr>
<th>Statement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC1. I need to understand the speaker’s implied meaning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC2. I must understand important details.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC3. I must have very good short memory.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC4. I must understand the context and topics of the conversation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC5. I must familiarize myself with CET4 pronunciation and intonations.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC6. I must infer the speakers’ attitude and views.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC7. I must catch important intonation and stresses of the speakers.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC8. I must understand the logic of the conversations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC9. I must make correct inferences of various inter-person relationships in the conversations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC10. a large vocabulary is important.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC11. background knowledge is important</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC12. I must understand the gist and catch the key words in dictation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC13. I must familiarize myself with the question type and find test taking skills for dictation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC14. I must familiarize myself with connected speech in English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What other skills/knowledge do you think are necessary to answer the listening comprehension questions in the CET-4? Please specify


<table>
<thead>
<tr>
<th>In order to do well in the Skimming and Scanning section, I must</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC1. skim the whole text to get the gist.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>RC2. look for specific words and locate the information precisely.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>RC3. have background knowledge</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>RC4. plan my time very well because time is a critical issue</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>RC5. find the topic sentence in each paragraph</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

| In order to do well in the questions in the Passage Reading section, |  |
|-------------------------------------------------------------|---|---|---|---|---|---|---|
| RC6. I must understand important details.                    | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| RC7. I must grasp the gist of the passages.                  | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| RC8. I must make inferences of implications.                 | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| RC9. background knowledge and common sense are important.    | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| RC10. I must use test-taking skills to eliminate wrong options. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

<table>
<thead>
<tr>
<th>In order to do well in the Banked Cloze section,</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC11. I must choose words based on understanding of its context.</td>
</tr>
<tr>
<td>RC12. I must use knowledge of prefix, suffix and roots to guess word meanings.</td>
</tr>
<tr>
<td>RC13. it is important to be familiar with test types and find out appropriate strategies.</td>
</tr>
<tr>
<td>RC14. I have to understand the relationships among different parts of the text.</td>
</tr>
</tbody>
</table>

What other skills or knowledge do you think are necessary to answer the reading comprehension questions in CET-4? Please specify.

| In order to do well in the Cloze section, |  |
|----------------------------------------|---|---|---|---|---|---|---|
| IS1. I must have knowledge of phrases and collocations. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| IS2. I must have knowledge of syntax and grammar. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| IS3. background knowledge or common sense are important. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| IS4. I must know the fine differences between synonyms. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| IS5. I must make inferences based on the context. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| IS6. It is necessary to understand the entire text. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| IS7. I must use test-taking skills to eliminate similar options. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

What other skills or knowledge do you think are necessary to answer the Cloze questions in the CET-4? Please specify.

219
Section B. Your perceptions of the CET-4 uses. This section includes statements about the purposes for you to take the CET-4, as well as the usefulness and values of the test to you. Please circle ONE number to indicate the extent to which you agree with each of the statements.

- **1 = Strongly disagree**  
- **2 = Disagree**  
- **3 = Somewhat disagree**  
- **4 = Neither agree or disagree**  
- **5 = Somewhat agree**  
- **6 = Agree**  
- **7 = Strongly agree**

**I take the CET-4 mainly to**  

<table>
<thead>
<tr>
<th>Purpose Statement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU1. be qualified for taking the CET-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU2. get the CET-4 Score Report for job seeking.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU3. prove my English proficiency.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU4. graduate with a degree.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU5. compete for academic awards and/or scholarship.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU6. prepare for postgraduate entrance exam.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU7. be qualified for CET4-SET.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU8. measure my English proficiency.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The CET score

PU9. provides an accurate measure of my English proficiency
PU10. provides an accurate measure of my English listening ability
PU11. provides an accurate measure of my English reading ability
PU12. provides an accurate measure of my English writing ability
PU13. is important for my studies in the university
PU14. is important for me to find a good job after graduation
PU15. provides useful feedback to my English language learning.
PU16. motivates me to work harder in English
Other values. Please specify.

Section C. Your expectation of taking the CET-4. This section includes statements about your self-efficacy in English proficiency and about your perceptions of the difficulty of the CET-4. Please circle ONE number to indicate the extent to which you agree with each of the statements.

• 1 = Strongly disagree • 2 = Disagree • 3 = Somewhat disagree
• 4 = Neither agree or disagree • 5 = Somewhat agree • 6 = Agree • 7 = Strongly agree

Considering the difficulty of CET-4 and my own ability, I have the confidence in doing well in
EX 1. the CET test
EX 2. short conversations in the listening comprehension section
EX 3. long conversations in the listening comprehension section
EX 4. passage listening in the listening comprehension section
EX 5. compound dictation in the listening comprehension section
EX 6. skimming and scanning in the reading comprehension section
EX 7. passage reading in the reading comprehension section
EX 8. banked cloze in the reading comprehension section
EX 9. translation
EX 10. cloze
EX 11. writing
Please indicate your perceptions of the difficulty of the different sections in the CET-4.

1= very easy, 2= easy, 3= somewhat easy, 4=neutral, 5= somewhat difficult, 6= difficult, 7=very difficult

12 13 14 15 16 17 18 19 20 21
EX 12. short conversations in the listening comprehension section ○ ○ ○ ○ ○ ○ ○
EX 13. long conversations in the listening comprehension section ○ ○ ○ ○ ○ ○ ○
EX 14. passages in the listening comprehension section ○ ○ ○ ○ ○ ○ ○
EX 15. compound dictation in the listening comprehension section ○ ○ ○ ○ ○ ○ ○
EX 16. skimming and scanning in the reading comprehension section ○ ○ ○ ○ ○ ○ ○
EX 17. Passage reading in the reading comprehension section ○ ○ ○ ○ ○ ○ ○
EX 18. banked cloze in the reading comprehension section ○ ○ ○ ○ ○ ○ ○
EX 19. Translation ○ ○ ○ ○ ○ ○ ○
EX 20. Cloze ○ ○ ○ ○ ○ ○ ○
EX 21. Writing ○ ○ ○ ○ ○ ○ ○

Section D. Your CET-4 preparation practice. This section includes statements about your CET-4 preparation practices. Please circle ONE number to indicate the extent to which you agree with each of the statements.

• 1 = Strongly disagree • 2 = Disagree • 3 = Somewhat disagree • 4 = Neither agree or disagree • 5 = Somewhat agree • 6 = Agree • 7 = Strongly agree

During CET preparation, I
TP1. spend more time on my weak points. ○ ○ ○ ○ ○ ○ ○
TP2. analyze CET4 question types to identify frequently assessed questions and tricky questions. ○ ○ ○ ○ ○ ○ ○
TP3. analyze CET-4 test papers to identify the level of difficulties for each section. ○ ○ ○ ○ ○ ○ ○
TP4. read CET-4 coaching books to know more about frequently assessed questions and test preparation strategies. ○ ○ ○ ○ ○ ○ ○
TP5. spend more time on those key questions that can be easily improved during test preparation. ○ ○ ○ ○ ○ ○ ○
### During CET preparation, I

<table>
<thead>
<tr>
<th>TP</th>
<th>Task Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP6</td>
<td>analyze CET4 score distribution to judge the relative importance of sections.</td>
</tr>
<tr>
<td>TP7</td>
<td>analyze my own performance so my test preparation can be more purposeful.</td>
</tr>
<tr>
<td>TP8</td>
<td>pay attention to the spelling when memorizing words.</td>
</tr>
<tr>
<td>TP9</td>
<td>memorize vocabulary including parts of speech, roots and prefixes.</td>
</tr>
<tr>
<td>TP10</td>
<td>memorize CET4 vocabulary or high-frequency words.</td>
</tr>
<tr>
<td>TP11</td>
<td>memorize linking words and phrases for writing</td>
</tr>
<tr>
<td>TP12</td>
<td>review the translation exercises in the textbook for translation.</td>
</tr>
<tr>
<td>TP13</td>
<td>repeatedly listen to CET-4 listening materials to familiarize myself with its pronunciation and intonation.</td>
</tr>
<tr>
<td>TP14</td>
<td>try to understand fully all the CET-4 listening materials I've practiced.</td>
</tr>
<tr>
<td>TP15</td>
<td>focus on understanding difficult and complex sentences in the passages I read.</td>
</tr>
<tr>
<td>TP16</td>
<td>practice composing essays using past CET-4 essay topics.</td>
</tr>
<tr>
<td>TP17</td>
<td>review the key sentence structures in the textbook for translation.</td>
</tr>
<tr>
<td>TP18</td>
<td>while listening, try to note down important information.</td>
</tr>
<tr>
<td>TP19</td>
<td>go over the options beforehand so as to focus my attention accordingly in listening.</td>
</tr>
<tr>
<td>TP20</td>
<td>while practicing reading, read questions before looking for key words and sentences in the passage.</td>
</tr>
<tr>
<td>TP21</td>
<td>while practicing reading, search for the answers in the text according to the sequence of questions.</td>
</tr>
<tr>
<td>TP22</td>
<td>while practicing writing, always try to use more advanced vocabulary.</td>
</tr>
<tr>
<td>TP23</td>
<td>while practicing writing, try to avoid grammar and spelling mistakes.</td>
</tr>
<tr>
<td>TP24</td>
<td>in practicing CET cloze, try to select answers out of the options by elimination strategies.</td>
</tr>
<tr>
<td>TP25</td>
<td>keep on practicing my spoken English.</td>
</tr>
</tbody>
</table>
During CET preparation, I

TP26. keep on reading English newspapers/websites. ○ ○ ○ ○ ○ ○ ○
TP27. keep on listening to English radio broadcasts. ○ ○ ○ ○ ○ ○ ○
TP28. keep on reading English aloud. ○ ○ ○ ○ ○ ○ ○
TP29. keep on writing diaries/blogs in English. ○ ○ ○ ○ ○ ○ ○
TP30. keep on communicating with English native speakers whenever possible. ○ ○ ○ ○ ○ ○ ○
TP31. keep on using English whenever possible, e.g. writing emails. ○ ○ ○ ○ ○ ○ ○
TP32. encourage myself to work hard on English. ○ ○ ○ ○ ○ ○ ○
TP33. try to learn from others. ○ ○ ○ ○ ○ ○ ○
TP34. try to build up my confidence in CET-4. ○ ○ ○ ○ ○ ○ ○
TP35. seek teachers’ advice on how to improve my CET-4 performance. ○ ○ ○ ○ ○ ○ ○
TP36. exchange my learning experience with classmates or friends. ○ ○ ○ ○ ○ ○ ○
TP37. reward myself whenever I made some progress. ○ ○ ○ ○ ○ ○ ○
TP38. consult senior students about the CET-4. ○ ○ ○ ○ ○ ○ ○
TP39. adjust my bio-clock so as to achieve my optimal state on the day when taking CET4. ○ ○ ○ ○ ○ ○ ○

What other things do you do to prepare for the CET-4? Please specify

Section E. About you

Your student No: ________________________________
Gender:  Male      Female
Your major _____________________________________
How long have you studied English? _____________ years.
Have you taken CET-4 before:   Yes       No
Are you familiar with the CET?
A. very familiar  B. familiar  C. somewhat familiar
D. not familiar  E. know nothing about it

End of the Questionnaire.

THANK YOU!
Appendix I. Excerpt from the CET Website about Score Interpretation

The following formula is used to convert the test-taker’s raw score to the reported standardized score:

\[
\text{TotSco} = \frac{(X - \text{Mean})}{\text{SD}} \times 70 + 500
\]

where \(\text{TotSco}\) is the total score; \(X\) represents the raw score of the test-taker before the conversion; \(\text{Mean}\) refers to the mean score of the norm group; \(\text{SD}\) is the standard deviation of the norm group.

There is a percentile position corresponding to each reported score as shown in the table below. For example, if a test-taker’s reported total score in the CET-4 is 450, his/her percentile position in the norm group is 25%, indicating that this test-taker’s score is higher than 25% of the norm group and lower than 75% of the group.

<p>| Reported Score and the Corresponding Percentile Position in the Norm Group |
|---------------|---------------|---------------|---------------|---------------|
| Listening comprehension (35%, total 249) | Reading comprehension (35%, total 249) | Integrated skills (10%, total 70) | Writing (20%, total 142) | Total (total 710) |</p>
<table>
<thead>
<tr>
<th>score</th>
<th>percentile</th>
<th>score</th>
<th>percentile</th>
<th>score</th>
<th>percentile</th>
<th>score</th>
<th>percentile</th>
<th>score</th>
<th>percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>100</td>
<td>1</td>
<td>25</td>
<td>1</td>
<td>55</td>
<td>1</td>
<td>330</td>
<td>1</td>
</tr>
<tr>
<td>110</td>
<td>2</td>
<td>110</td>
<td>2</td>
<td>28</td>
<td>2</td>
<td>60</td>
<td>2</td>
<td>350</td>
<td>2</td>
</tr>
<tr>
<td>120</td>
<td>3</td>
<td>120</td>
<td>4</td>
<td>31</td>
<td>3</td>
<td>65</td>
<td>3</td>
<td>370</td>
<td>4</td>
</tr>
<tr>
<td>130</td>
<td>7</td>
<td>130</td>
<td>9</td>
<td>34</td>
<td>5</td>
<td>70</td>
<td>5</td>
<td>390</td>
<td>7</td>
</tr>
<tr>
<td>140</td>
<td>12</td>
<td>140</td>
<td>17</td>
<td>37</td>
<td>7</td>
<td>75</td>
<td>9</td>
<td>410</td>
<td>11</td>
</tr>
<tr>
<td>150</td>
<td>18</td>
<td>150</td>
<td>26</td>
<td>40</td>
<td>12</td>
<td>80</td>
<td>16</td>
<td>430</td>
<td>17</td>
</tr>
<tr>
<td>160</td>
<td>26</td>
<td>160</td>
<td>37</td>
<td>43</td>
<td>17</td>
<td>85</td>
<td>25</td>
<td>450</td>
<td>25</td>
</tr>
<tr>
<td>170</td>
<td>37</td>
<td>170</td>
<td>48</td>
<td>46</td>
<td>28</td>
<td>90</td>
<td>37</td>
<td>470</td>
<td>34</td>
</tr>
<tr>
<td>180</td>
<td>48</td>
<td>180</td>
<td>62</td>
<td>49</td>
<td>39</td>
<td>95</td>
<td>52</td>
<td>490</td>
<td>44</td>
</tr>
<tr>
<td>190</td>
<td>59</td>
<td>190</td>
<td>72</td>
<td>52</td>
<td>51</td>
<td>100</td>
<td>66</td>
<td>510</td>
<td>55</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>200</td>
<td>82</td>
<td>55</td>
<td>63</td>
<td>105</td>
<td>77</td>
<td>530</td>
<td>65</td>
</tr>
<tr>
<td>210</td>
<td>82</td>
<td>210</td>
<td>90</td>
<td>58</td>
<td>78</td>
<td>110</td>
<td>86</td>
<td>550</td>
<td>74</td>
</tr>
<tr>
<td>220</td>
<td>91</td>
<td>220</td>
<td>96</td>
<td>61</td>
<td>90</td>
<td>115</td>
<td>92</td>
<td>570</td>
<td>83</td>
</tr>
<tr>
<td>230</td>
<td>97</td>
<td>230</td>
<td>98</td>
<td>64</td>
<td>97</td>
<td>120</td>
<td>97</td>
<td>590</td>
<td>90</td>
</tr>
<tr>
<td>240</td>
<td>99</td>
<td>240</td>
<td>99</td>
<td>67</td>
<td>99</td>
<td>125</td>
<td>99</td>
<td>610</td>
<td>95</td>
</tr>
</tbody>
</table>

\[\text{Percentile} = \left(\frac{\text{Scaled Score} - \text{Mean}}{\text{SD}}\right) \times 70 + 500\]
Appendix J. Covariance Matrix for the Construct Indicators in the CET-4 Washback Model

<table>
<thead>
<tr>
<th></th>
<th>1.06</th>
<th>0.68</th>
<th>0.70</th>
<th>0.78</th>
<th>0.55</th>
<th>0.73</th>
<th>0.79</th>
<th>0.68</th>
<th>0.38</th>
<th>0.37</th>
<th>0.66</th>
<th>0.64</th>
<th>0.65</th>
<th>-0.07</th>
<th>0.63</th>
<th>0.22</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.68</td>
<td>1.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.79</td>
<td>1.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.78</td>
<td>0.75</td>
<td>0.81</td>
<td>1.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td>0.55</td>
<td>0.78</td>
<td>0.58</td>
<td>2.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.73</td>
<td>0.71</td>
<td>0.70</td>
<td>0.75</td>
<td>1.04</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.79</td>
<td>0.74</td>
<td>0.70</td>
<td>0.77</td>
<td>1.30</td>
<td>1.15</td>
<td>1.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.68</td>
<td>0.66</td>
<td>0.58</td>
<td>0.65</td>
<td>0.76</td>
<td>0.78</td>
<td>0.94</td>
<td>1.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.38</td>
<td>0.40</td>
<td>0.62</td>
<td>0.40</td>
<td>0.57</td>
<td>0.64</td>
<td>0.52</td>
<td>0.31</td>
<td>1.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.37</td>
<td>0.46</td>
<td>0.60</td>
<td>0.43</td>
<td>0.91</td>
<td>0.75</td>
<td>0.84</td>
<td>0.74</td>
<td>0.53</td>
<td>2.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.66</td>
<td>0.69</td>
<td>0.64</td>
<td>0.69</td>
<td>0.87</td>
<td>0.84</td>
<td>1.01</td>
<td>0.84</td>
<td>0.43</td>
<td>1.23</td>
<td>1.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.64</td>
<td>0.69</td>
<td>0.63</td>
<td>0.67</td>
<td>0.79</td>
<td>0.79</td>
<td>0.90</td>
<td>0.78</td>
<td>0.46</td>
<td>0.94</td>
<td>0.85</td>
<td>1.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.65</td>
<td>0.68</td>
<td>0.60</td>
<td>0.65</td>
<td>0.84</td>
<td>0.83</td>
<td>1.06</td>
<td>0.86</td>
<td>0.42</td>
<td>1.23</td>
<td>1.00</td>
<td>0.86</td>
<td>1.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.42</td>
<td>-0.50</td>
<td>-0.02</td>
<td>-0.17</td>
<td>0.63</td>
<td>0.45</td>
<td>-0.97</td>
<td>0.09</td>
<td>0.02</td>
<td>-0.55</td>
<td>0.03</td>
<td>41.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.63</td>
<td>0.66</td>
<td>0.30</td>
<td>0.40</td>
<td>0.62</td>
<td>0.53</td>
<td>0.99</td>
<td>0.92</td>
<td>-0.30</td>
<td>0.83</td>
<td>0.66</td>
<td>0.22</td>
<td>0.58</td>
<td>23.76</td>
<td>45.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.22</td>
<td>0.28</td>
<td>-0.09</td>
<td>0.14</td>
<td>0.62</td>
<td>0.05</td>
<td>0.48</td>
<td>0.49</td>
<td>-0.36</td>
<td>-0.02</td>
<td>0.21</td>
<td>-0.19</td>
<td>0.29</td>
<td>13.88</td>
<td>13.46</td>
<td>15.85</td>
<td></td>
</tr>
</tbody>
</table>