The Hong Kong Polytechnic University

Department of English

Applying an Assessment Use Argument to Investigate a College-Level English Language Test in Universities in Xi'an

LIU Min

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

February 2013

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to be best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____________________ (Signed)

_____LIU Min________ (Name of Student)

ABSTRACT

This study applied an Assessment Use Argument (AUA) to investigate a college-level English language test in universities in Xi'an, namely, the College English test Band Four (CET-4). The overarching research purpose is to investigate to what extent and in what way the CET-4 can serve as a useful indicator of students' overall English proficiency and an effective measure to promote College English teaching and learning. The related research questions were first linked to the corresponding claims on interpretations of scores, decisions made on scores, and consequences of test uses in an AUA. Then the study narrowed down its research foci: 1) to examine the construct of the CET-4 and its content relevance and coverage, 2) to identify factors underlying the multiple decisions made on CET-4 scores, 3) to reveal stakeholders' perceptions of the CET-4 and its washback, and explore the possible relationships between students' perceptions and their test performances.

A distinctive feature of this study is the articulation of an AUA for the CET-4 within China's EFL assessment context. There has been thus far a dearth of research that draws on the structure and rationale of an AUA to either develop a test or justify test uses. Therefore, this study offers an exemplary attempt to examine consequences of the CET-4 while weighing the validity of the revised listening and reading components. The AUA offers an overarching logical structure and a conceptual guidance to investigate all the research questions and sub-questions, and the corresponding claims and warrants.

A mixed-method approach was employed to collect backing evidence. About 900 students and 200 teachers participated in this study. A quantitative approach was adopted to analyze the large volume of test data and questionnaire surveys, while a qualitative approach was applied to the analyses of test contents and interview data. Evidence from multiple sources was triangulated to strengthen the logic and coherence of the AUA for the CET-4.

Data were analyzed in two phases. In the preliminary study, a statistical comparative study was conducted with 188 test takers' valid scores from the old and the new versions of the CET-4. In the main study, correlations and exploratory factor analysis were performed on a larger pool of 2692 valid score cases from the CET-4. Results from both studies evidence that the current CET-4 possesses better internal structure. In addition, the listening and reading components of seven authentic CET-4 papers underwent content analysis with five parameters including text length, readability, genres, topics, and skills coverage. The results demonstrate an overall nice correspondence between test contents and descriptions on characteristics of input and characteristics of expected response in uniform teaching and testing syllabuses, indicating that the revised listening and reading components have a higher degree of content validity. Questionnaires and interviews, employed to investigate decisions made on CET-4 scores and explore the underlying factors in these decision-making processes, reveal that using test scores as a gatekeeper in selection, advancement, or competition takes a deep root in the inherent influences of China's imperial examination system. Institutional decisions manifest a tendency of using large-scale and high-stakes tests as a catalyst or a lever for curriculum innovation. Both test designers and test users should be held accountable for stakeholders to be affected by their decisions. In addition, interview and survey data explore stakeholders' perceptions of test design, test influences, teaching and learning practices, test preparation activities and so on. Multiple regression analyses reveal that the students' motivations, their perceived difficulty factors, and test-taking strategies have influences on their test performances.

To sum up, the study reveals a multiplicity of perspectives on the conceptions, analyses and arguments that bear on assessment validity, use and consequences. The study provides the CET stakeholders, especially test users and test developers, with useful insights to help deepen the understanding of the concept of test validity within the framework of an AUA and also shed light on the process of assessment justification in the Chinese EFL context.

RESEARCH OUTPUT ARISING FROM THE THESIS

Book Chapters

Liu, M., & Qian, D.D. (in press). Investigating the legitimacy of decisions based on CET-4 scores: Applying an assessment use argument approach. In D. D. Qian & L. Li (Eds.), *Teaching and learning English in East Asian Universities: Global Visions and Local Practices*. UK: Cambridge Scholars Publishing.

Liu, M., & Qian, D.D. (2010). Evaluating the new College English Test Band 4: A preliminary investigation. In L. Li & D. D. Qian (Eds.), *English language education in Asian universities: Classroom practices and research issues* (pp.12-22). Hong Kong: The Hong Kong Polytechnic University.

Conference Presentations

Liu M. (2012, August). *Building an assessment use argument for the College English Test Band Four*. Paper presented at the 8th International Symposium on Teaching English at Tertiary Level & The 17th International Conference of Pan-pacific Association of Applied Linguistics (PAAL), Beijing, China.

Liu M. (2011, November). *Justifying the test use of the College English Test Band Four with an assessment use argument*. Paper presented at the 2nd combined conference of ALAA-ALANZ of Canberra Langfest, Canberra, Australia.

Liu M. (2011, August). *An investigation into the construct validity of the CET-4 with an assessment use argument*. Paper presented at the 16th AILA2011, the World Congress of Applied Linguistics, Beijing, China.

Liu M. (2011, August). *An investigation into the test use of the College English Test Band Four with an assessment use argument*. Paper presented at the 7th KELTA: 2011 International Conference, Korea English Language Testing Association, Seoul, South Korea.

Liu M. (2010, July). *Evaluating the impact of the College English Test Band Four*. Paper presented at the 2nd APACLSP Conference, the Asia-Pacific Rim LSP and Professional Communication Association, Kuala Lumpur, Malaysia.

Liu M. (2010, April). *A construct validation of the listening and the reading components of the new CET-4 with an assessment use argument*. Paper presented at the 3rd Postgraduate Research Symposium: Language and Culture Studies in the Pearl River Delta, Hong Kong.

Liu M. & Qian, D.D. (2010, April). *A preliminary investigation into the new CET-4 with a comparative approach*. Poster presented at LTRC 2010, Language Testing Research Colloquium, Cambridge, UK.

Liu M., & Qian, D.D. (2009, December). *An exploration into the new College English Test Band Four: A comparative study*. Paper presented at the 3rd HAAL Conference, Hong Kong Association for Applied Linguistics, Hong Kong.

Liu M., & Qian, D.D. (2009, November). *A preliminary exploration into the social and educational impact of the College English Test Band Four*. Paper presented at Symposium on English as the Language of Asian Business and Professions, Research Centre for Professional Communication in English, Hong Kong.

Liu M.,& Qian, D.D. (2009, October). *Evaluating the new College English Test Band 4: A preliminary investigation*. Paper presented at the 5th International Symposium on Teaching English at Tertiary Level, Hong Kong.

Liu M., & Qian, D.D. (2009, March). *An empirical study on the washback effect of the old CET-4 — from the perspective of communicative competence*. Paper presented at the 2009 Annual PolyU Faculty of Humanities Postgraduate Research Symposium, Hong Kong.

ACKNOWLEDGEMENTS

The completion of this study would not have been possible without support and participations of many people.

First and foremost, I would like to express my deepest gratitude to my doctoral supervisor, Prof. David Qian, for his insightful advice and boundless support through my PhD study. He has always demonstrated his professional expertise in language assessment field and offered his invaluable comments, patience, and encouragement during the writing and revision of this dissertation.

I would like to express my thanks to Prof. Martin Warren, my co-supervisor. I still remember our meeting on my first day to PolyU. Martin patiently explained the procedures a new student should follow and gave me an overall guidance on how to plan my PhD study strategically. I am also grateful to Dr. Chris Green. His methodology course helped me broaden my knowledge and perspectives on research design. Chris is also the examiner in my confirmation report, offering me constructive suggestions at the early stage of my PhD study. My sincere thanks go to Dr. Stephen Evans for his informative workshops and mentoring in EEPRS program. His suggestions and revisions helped improve quality of my dissertation. I would like to extend my thanks to faculty of Department of English at The Hong Kong Polytechnic University. I have benefited a lot from their professional teaching and enlightening lectures.

I am much obliged to Prof. Lyle Bachman. During my study at UCLA, I benefited a lot from his instructions and our discussions about the Assessment Use Argument framework. The inspirations from his framework helped shape the theoretical underpinnings of this study. I am sincerely thankful for Prof. Bachman’s comments on and the revisions of my earlier draft.

A special acknowledgement also goes to Prof. Wu Shinian at the Grand Valley State University. During my MA study, I took his course Second Language Assessment, which stimulated my initial interest in language assessment. Over the years, Prof. Wu has always been a good mentor and advisor when I inquired and exchanged ideas with him on assessment issues during my dissertation research.

I am particularly grateful to my Board of Examiners, Prof. Antony Kunnan, Dr. Lu Danhuai, and the Chair of BoE, Dr. Li Lan, for their careful reading, prompt feedback and illuminating suggestions.

I wish to express my thanks to all the participants from the four sampled universities in Xi’an for their kind cooperation and valuable information. Without their help, this empirical study would not have been accomplished.

Last but not least, I am profoundly indebted to my family for their unconditional love, understanding, and dedication over the years. A special thank is to my beloved son. When I started my PhD study at PolyU, he was only one and half years old. I am profoundly indebted to him.

My sincere appreciation goes to all of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AERA	American Education Research Association
ANOVA	Analysis of Variance
APA	American Psychological Association
AUA	Assessment Use Argument
Cambridge ESOL	The University of Cambridge English for Speakers of Other Languages
CECR	College English Curriculum Requirements
CET-4	College English Test Band Four
CET-6	College English Test Band Six
CFA	Confirmatory Factor analysis
DIF	Differential item Functioning
EAP	English for Academic Purpose
ECD	Evidence Centered design
EFA	Exploratory Factor analysis
EFL	English as a Foreign Language
ESL	English as a Second Language
ETS	Educational Testing service
GSEEE	Graduate School Entrance English Examination
HED	Higher Education Department
IELTS	International English Language Testing System
MCQ	Multiple Choice Questions
MoE	Ministry of Education
NCETC	National College English Test Committee
NCME	National Council on measurement in Education
NMET	National Matriculation English Test
P/MEC	Provincial or Municipal Education Commission
SAQ	Short Answer Questions
SEM	Structural Equation Modeling

SET	Spoken English Test
SQ	Student Questionnaire
SRF	Score Report Form
TOEFL iBT	Test of English as a Foreign Language: Internet-based Test
TQ	Teacher Questionnaire

CHAPTER 1

INTRODUCTION

1.1 Statement of the problem

The scope and conceptions of validity have evolved over the past decades. In the early discussions, validity had been viewed as a componential concept (e.g., Anastasi, 1954; Angoff, 1988; Cronbach & Meehl, 1955; Cureton, 1951; Lado, 1961). The 1980s witnessed concept syntheses of validity. Messick (1989) in his seminal article defined validity as a unitary and multifaceted concept, encompassing value implications and social consequences. Since then, Messick's four-fold progressive validity matrix has drawn language testers' attention to score interpretation as well as test use.

Inspired by Messick's unified model of validity, quite a number of researchers have investigated the consequences of test uses, in particular the washback of high-stakes tests (e.g., Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Andrews, 1994; Cheng,1997, 1998; Green, 2007; Messick, 1996; Qi, 2004; Schohamy, 1993, 2001; Wall, 1996; Watanable, 1996). These studies have broadened and deepened our understanding of washback as a concept and a complex mechanism. However, in spite of rich findings on scope and intensity of washback, washback was treated as unrelated to validity in these discussions.

Other researchers have made theoretical and empirical attempts to link validity to test uses, consequences and ethical considerations. Hamp-Lyons (1997) listed washback to a broader concept of impact and addressed washback and validity from ethical concerns in language testing. Bachman and Palmer (1996) listed six qualities of test usefulness and incorporated validity and impact into a unitary concept of test usefulness. Lynch (2001) drawing on postmodern critical theory integrated validity with ethical considerations in his five categories of validity framework. Kunnan (2000, 2004) proposed test fairness framework that not only links validity and consequences but also introduced three new qualities: absence of bias, access and administration. However, “while these studies enlarged our perspective beyond Messick’s unitary validity model, they failed to provide an explicit link between validity and test use. The test qualities they have articulated have no clear logical mechanism for integrating these into a set of procedures for test developers and users to follow” (Bachman, 2005, p.7).

Recent validity studies have elaborated on argument-based approaches to validation. Mislevy (1996) and colleagues (Mislevy, Steignberg & Almond, 2002, 2003) developed an approach called evidence centered design (ECD). It provides detailed steps and procedures in stages of validation, logically integrating construct definitions, characteristics of assessment tasks, and psychometric models needed to deal with complex performance data (Bachman, 2005). Kane (1992, 2004, 2006) and colleagues (Kane, Crooks, & Cohen, 1999) developed a notion of interpretative argument for linking observations to interpretations, and generalizing from evidence of test performance to inferences to be made on test takers. A distinctive emphasis in Kane’s work is the need to systematically anticipate threats to the validity of these inferences (McNamara, 2006). Their discussions provide a series of logical inferences to investigate score-based interpretations, but do not address in detail the facets of values, cultures, and consequences raised in Messick’s framework. Mislevy ignored the issue of the consequences of test use, while Kane, in spite of embracing a concern for consequences, did not develop a methodology to investigate them (Bachman, 2005; McNamara, 2006).

Given the lack of a clear linkage from test performances to interpretations of scores, and to test uses and consequences of test uses, Bachman, drawing on Toulmin’s argument model and Messick’s unitary and multifaceted validity framework, proposed an Assessment Use Argument (AUA, Bachman, 2003, 2005; Bachman & Palmer, 2010). The AUA derives its structure from Toulmin’s model, including elements of claims, warrants, backing and rebuttals. Bachman advanced the argument-based approach by embedding test use and consequences that

Messick proposed to incorporate in the validation framework into his utilization argument, which along with the validity argument constitutes an Assessment Use Argument. Bachman and Palmer's AUA offers an overarching logic structure and a conceptual guidance for an explicit and coherent linkage from test performance to interpretations and from interpretations to uses. An AUA not only serves as a framework to guide test development and test use, but also provides a basis for test developers and decision makers to be held accountable to stakeholders to be affected by the use of the assessment and the decisions that are made on it.

So far, there has been a dearth of empirical research adopting an AUA to either develop a test or justify test use. Thus, the present study is intended to apply the AUA framework to investigate a college-level English language test in China. More specifically, the study aims to offer an exemplary attempt to link test interpretations with test consequences by employing the AUA in the Chinese higher education assessment context. This college-level English language test is addressed as the College English Test Band Four (CET-4[1]). The following section will further discuss the research motivations on why this study is situated within the setting of the CET-4 to better contextualize the research background.

1.2 Context of the study

The history of examination in China can be traced back to Han Dynasty. Its educational system has been characterized as examination-oriented system (Cheng, 2008, 2010; Qi, 2007; Spolsky, 1995). Today testing and examinations maintain their important and powerful roles in educational system (Li, 1990). Students need to take numerous examinations throughout their schooling. Currently three main large-scale examinations are being implemented at senior secondary, tertiary and postgraduate levels of education, namely, the National Matriculation English Test (NMET), the College English Test Band Four (CET-4), and the Graduate School

[1] This study collects test-takers' scores from official CET-4 administrations, scores from a university's final English examination administered to sophomores, and seven actually used CET-4 test papers published by a commercial publishing house. Hence, for conciseness the CET-4 is used hereinafter referring to this college-level English language test as a whole.

Entrance English Examination (GSEEE). One point needs to be clarified here is that the College English Test Band Six (CET-6) is also administered at the tertiary level. However, only students whose CET-4 scores are beyond 425 are eligible to take it. In addition, the CET-6 is optional to eligible potential candidates, who can make their own decisions on whether to take the CET-6. Hence, the present study is mainly focused on the CET-4 in its discussion of the college-level test.

The above three tests are mentioned herein because they embody some typical test features in English as a foreign language test (EFL) assessment context in China. They are recognized as large-scale due to their vast test population. Take the year of 2007 for example, the annual candidatures for the NMET, CET, and the GSEEE were 10.1 million, 12.5 million, and 1.2 million respectively (He, 2010; Jin, 2010; Qi, 2010). In addition, they are also acknowledged as high-stakes tests due to the fact that performances on these tests will weigh heavily on students' chances of admission and graduation. In the following part, the major uses of these three tests are discussed first to facilitate our understanding of features of high-stakes tests. The multiple uses of the CET-4 stand out in comparison with other two tests. The section concludes with major research motivations on how the present study is shaped.

1.2.1 Multiple uses of the CET-4

The most fundamental and prevalent use of language tests is to provide information for making decisions about individuals and programs (Bachman, 1990). The decisions based on language tests may apply in both educational and social situations. Decisions made on test scores within educational settings are usually related to admission, placement, graduation and curriculum reform, while decisions made in social dimensions are pertinent to employment, promotion, certification, and immigration. These decisions in turn will have consequences on individuals, programs, instructors, organizations or societies (Bachman & Palmer, 2010).

The major use of both the NMET and the GSEEE is to make decisions on selection and admission of students whose English proficiency has reached a required level to qualify them for the future undergraduate and postgraduate studies. With regard to the CET-4, it is necessary to examine its intended purposes first. The CET-4 Syllabus stipulates that the College English Test is aimed at measuring precisely college students' comprehensive employment of English and playing an active role in realizing the objectives of college English teaching (National College English Testing Committee, 2006). The National College English Testing Committee (NCETC) has reiterated in a series of publications that the CET-4 is to provide an objective evaluation of a student's overall English proficiency and positively impact EFL teaching at the tertiary level in China (e.g., Jin, 2005, 2006, 2008; Jin & Yang, 2006; Yang & Weir, 1998). It is posted on the official CET website (www.cet.edu.cn) that the College English Test is designed as an objective and accurate evaluation of the English proficiency of the college students in order to better inform the English teaching of non-English majors in the institutions of higher learning in China (CET, 2011). However, the CET-4 in reality has been used to serve multiple uses. The NCETC admits that test results have been used to make decisions on graduation, employment, and even a residence permit in some major cities (Jin, 2005, 2008). Just as Shohamy (2001) comments, "doing well on tests can take a person to the best university and open the way to an excellent education; doing poorly can send a person to a low level university and block the possibility of higher education"(p.15). The high-stakes nature of these tests is embodied in such life-changing decisions.

Multiple uses of an assessment can also be observed in large-scale and high-stakes international tests like Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS). TOEFL is used for making decisions related to university admission, scholarship, and placement into institutional English as a second language (ESL) courses (ETS, 2010). The academic module of IIELTS is used for decisions related to institutions of higher and further education while the general training module is used for employment, immigration, and professional accreditation (IELTS, 2010). Likewise, the use of the CET-4 in China is not restricted to tertiary education, as discussed earlier. Regardless of their decisions and uses, it can be seen that large-scale and high-stakes English tests tend to be used as gatekeepers to admission, selection, employment and promotion (Cheng, 2008, 2010; Qi, 2004, 2007).

The above discussion explains why this study is targeted at the CET-4 rather than the NMET or the GSEEE in spite of their shared large-scale and high-stakes attributes in Chinese EFL assessment context. The multiple uses of the CET-4 make the test more salient and noteworthy for investigation. The present study is focused on test uses by different groups of stakeholders as well as the underlying factors in test uses. The study will cover investigation into decisions made on CET-4 scores, decision-makers to be held accountable, and evidence to support the decisions and justify the test uses.

1.2.2 Washback of the CET-4

There has been a long controversy on washback of the CET-4 on College English teaching and learning. Test designers and test administrators maintain that the test has positive effects on English teaching at the overall tertiary level. Its implementation has strongly motivated teachers and students to attach more importance to English study. In addition, it has met social needs and gained social recognition (Jin, 2006, 2008; Jin & Yang, 2006; Wu, 2005; Yang & Weir, 1998).

At the same time, the CET-4 has invited criticisms from language educators and testers. Some accused the test of leading to the phenomena of “teaching and learning to the test” and “higher ability, lower proficiency” (Gu, 2007; Liu & Dai, 2004), indicating that even though some students could obtain higher CET-4 scores, they were still unable to communicate in English effectively. Some censured the hidden policy of linking students’ CET-4 performances to their Bachelor’s degrees (Cai, 2005, 2006; Gu, 2003). Even test designers recognized some of its negative effects such as narrowing down teaching curriculum, and replacing textbooks with coaching materials (Jin, 2006, 2008). The CET-4 has thus become the most debated and controversial test in China (Cheng, 2008).

A consensus has been reached in the field of language testing that a test has effects on teaching and learning, namely, washback. In particular, it has been acknowledged that washback of a large-scale and high-stakes test is inevitable (e.g., Alderson & Wall, 1993; Cheng, 1997, 2008; Qi, 2003; Shohamy, 2001; Wall, 2005; Watanable, 1996). However, washback is more than identifying stakeholders to be affected and summarizing the observed effects of an assessment on teaching and learning. Washback mechanisms and the myriad factors underlying washback phenomenon are more complex than assumed. Empirical research is thus called for to support any claims on washback of an assessment, regardless of its being intended or unintended, positive or negative. This trend in the field of language assessment serves as one of the research motivations.

1.2.3 The CET-4 reform in 2006

The NCETC has been making painstaking efforts to improve the CET-4. In order to mitigate negative washback of the CET-4, and meet social demands for graduates with higher English communicative competence, the Ministry of Education (MoE) initiated a series of reforms. *The College English Curriculum Requirements* (hereinafter the CECR) promulgated in 2004 and the CET-4 Syllabus revised in 2006 unveiled the reform on the College English Test. With the purpose to improve test validity and to maximize its positive washback on College English teaching, the CET-4 underwent its largest reform since its inception in 1987. In December 2006, the NCETC launched the revised CET-4 (alternatively called the new CET-4 by the public). Since the teaching objective lays more emphasis on cultivating students' use of English in an all-around way, particularly their listening abilities, the CET-4 has made the significant changes in the increase of Listening weight (from 20% to 35%) and the inclusion of Long Conversations and Compound Dictation sections. In addition, sections of

Skimming and Scanning, and Banked Cloze are newly incorporated into the test while the Vocabulary and Structure section is cancelled. In terms of the test format, more constructed response items are adopted to reduce the dominance of traditional selective response items.

The launch of the current CET-4 provides a natural and appropriate impetus to conduct an evaluation of the revised version, since validation is a never-ending process that involves accumulating various evidence to support score interpretations (Bachman, 1990; Cronbach, 1988; Messick, 1989, 1996). As long as the efforts to improve the test quality continue, there is a continuing need to validate the CET-4. In addition, since the current CET-4 was launched in 2006, thorough and profound empirical research on the new version has not been sufficiently conducted. This stimulated the present study to cover questions like: to what extent these newly-adopted test tasks can measure what they are intended to measure, how stakeholders perceive and evaluate the revised test, and whether the new elements in the current CET-4 can bring about particular changes expected by test designers.

To sum up, the aforementioned discussion delineated how the study was prompted and shaped by current trends in the field of language testing, and the status quo of the CET-4 against the background of Chinese EFL assessment contexts. Thus, the present study endeavors to articulate a specific AUA for the CET-4. The validity argument in the AUA framework links test takers' assessment performances on the CET-4 to interpretations of their English proficiency to weigh the validity of the test. Its utilization argument links score-based decisions to consequences of the test and of the decisions to investigate the test uses.

1.3 Research questions

In response to the intended purposes of the CET-4 and the objective of its reform in 2006, the overarching research purpose of the present study is to investigate to what extent and in what ways the CET-4 can serve as a useful indicator of students' overall English proficiency and an effective measure to promote College English teaching and learning. In other words, the study is aimed at investigating the extent to which CET-4 scores can be interpreted as the English proficiency defined in the corresponding teaching and testing syllabuses, and the extent to which the CET-4 has fulfilled the intended washback envisaged by the 2006 reform. Accordingly, three research questions (RQ) are proposed as follows:

RQ1: To what extent can the CET-4 serve as an indicator of students' English proficiency?

RQ2: What evidence has been provided or is needed to justify the major types of decisions made based on CET-4 scores?

RQ3: In what ways and to what extent can the CET-4 and the decisions made based on it affect English teaching and learning?

The first research question seeks to explore the validity of the test. The second research question investigates multiple uses of the test and serves as an essential bridge to link interpretations of scores to consequences of test use. The third research question actually explores the effects of the CET-4 on English learning and teaching activities, namely washback. The investigation of the CET-4 will be primarily focused on the constructs of the reformed listening and reading components. For one thing, the most significant revisions are made in the contents and formats of the two subtests. For another, the two subtests altogether occupy 70% of the total weight. The consequences of using the test and of decisions made based on its scores will be examined in terms of the overall test.

The above three research questions will be further discussed in Chapter 4. They will be linked to the corresponding claims of the AUA for the CET-4 first (see section 4.2). Next, more focused and operationally defined sub-questions are to be elaborated with the articulation of corresponding warrants or rebuttals (see section 4.4). In this way, the AUA for the CET-4 can serve as a conceptual framework to guide the investigation of all the research questions.

1.4 Overview of the research design

To better answer the research questions and to ensure the validity of the study, a mixed-method of both qualitative and quantitative research design is employed, mainly consisting of test content analysis, interviews, document analysis, test scores, and questionnaire surveys, so that evidence can be collected from multiple sources to achieve data and instrument triangulation (see Figure 1.1).

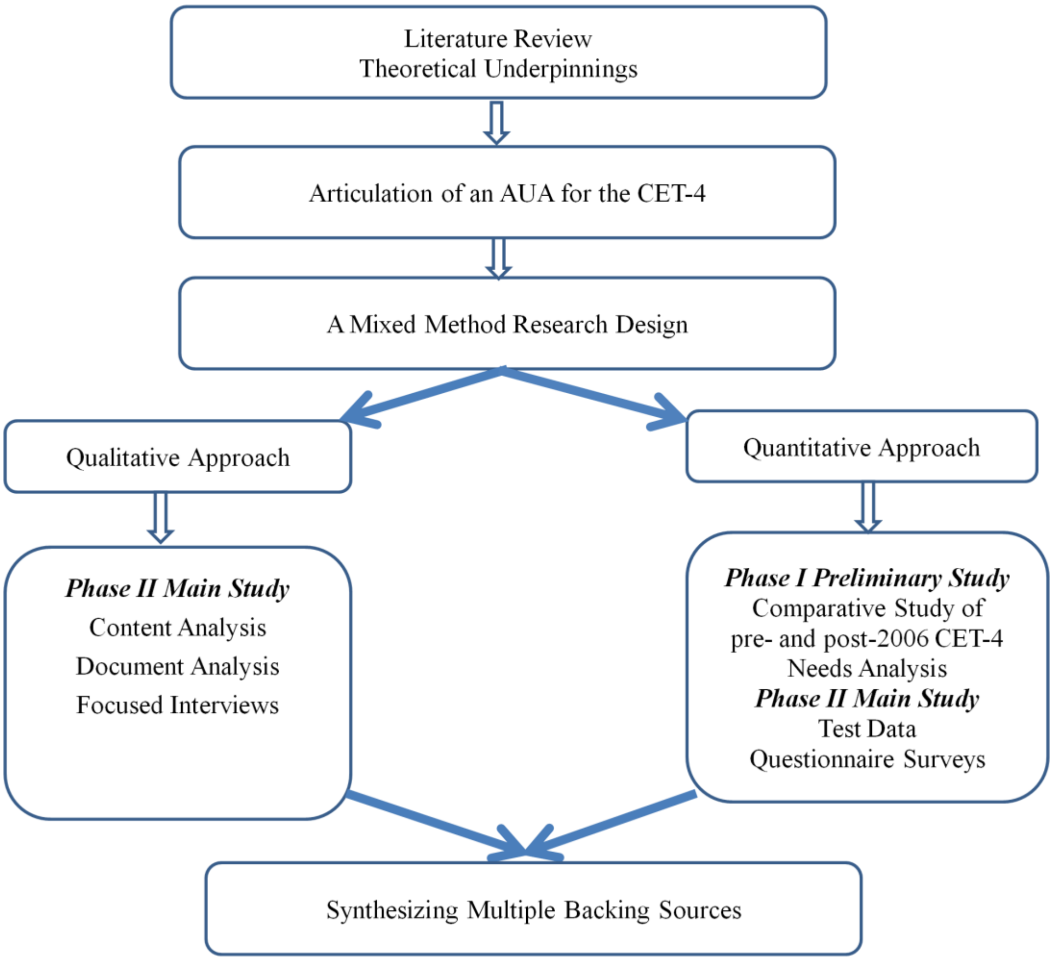

Figure 1.1 Flowchart of research methodology

1.5 Significance of the study

The investigation into the CET-4 has a number of far-reaching implications which are expected to make theoretical, methodological and pedagogical contributions to the field of language assessment.

Theoretically, the present study is expected to provide empirical evidence to help us gain a broader view and more profound understanding of the concept of validity within the framework of an AUA and the process of assessment justification. In particular, by addressing validity of the CET-4 and its washabck, it is hoped that the study can contribute to our understanding of linking validity issues to washabck mechanism within an AUA framework.

From a methodological perspective, first, the present study employs current thinking in educational measurement and language assessment by drawing on Bachman and Palmer's (2010) AUA, which provides a logical and coherent link and the rationale needed to justify the interpretations and uses based on test takers' performances. Second, the major instruments adopted in washabck studies are questionnaire surveys, interviews and classroom observations to investigate relationship between testing, teaching and learning. Few studies have linked test takers' attitudes and behaviours revealed from their questionnaire survey responses to their test performances. Questionnaires in this study not only underwent basic descriptive analyses of frequency counts, means, correlations, but also underwent inferential analyses of independent t-test and multiple regression analyses. The study is expected to make a methodological contribution to argument-based validation literature by operationalizing Bachman and Palmer's AUA, and to washback research by combining a traditional qualitative approach with inferential quantitative analyses in order to yield findings more meaningful, interpretable and illuminating.

As far as the pedagogical perspective is concerned, testing and teaching are closely related (Heaton, 1988). While this study links investigation on validity of the

CET-4 and consequences of its uses together, in an attempt to discover its merits and demerits and further improve test usefulness, the outcome of this analysis will lead to some suggestions to college English teaching, learning and testing. Furthermore, the notion of Communicative Language Ability and the communicative teaching approach have been widely acknowledged and implemented, which accordingly calls for the practice of a communicative language testing. The task of the CET designers is to provide a comprehensive assessment of test takers' communicative language ability (Jin, 2006). Thus, this study has the potential to promote the practice of communicative testing. To maximize the positive washback on College English teaching and learning is also one of the major purposes to reform the CET-4. Hence, it is hoped that by analyzing the CET-4 and investigating the consequences of its uses and of the decisions that are made on it, more light can be shed on the complexity and mechanism of washback in terms of promoting the communication-oriented teaching and learning practices.

In brief, the study is expected to reveal a multiplicity of perspectives on the conceptions, analyses and arguments that bear on validity, test uses and consequences in language assessment. Moreover, it will provide information for different groups of stakeholders mainly including test designers, EFL program administrators, teachers and test takers, and contribute to test development, validation, teaching and learning practices.

1.6 Thesis organization

The thesis consists of ten chapters. Chapter 1 introduces the study in terms of research context, research questions, methodology approach, and the thesis organization. The purpose is to present an overview of the background to and the rationale of the study so as to inform readers why the study targets the CET-4 within China's EFL assessment context, what prompts the study to draw on an AUA, and what is expected to achieve in this study.

Chapter 2 traces the evolution of the CET-4 from its inception in 1987 through to its largest reform in 2006. It first introduces the test purpose, candidature, and organization of the CET-4. Next, its development and test revisions before 2006 are reviewed, along with characteristics of College English teaching and major trends in language assessment field at the corresponding stage. It then emphasizes the current CET-4, delineating its reform background and introducing major features of its test contents and formats. The chapter concludes with reviews on domestic validation and washback studies on the CET-4.

Chapter 3 examines the theoretical underpinnings that guide the present study. It first reviews the evolution of validity and validation approaches by tracing its origin, disputes on its definition and scope, and recent argument-based approaches to validation. Following this, it provides an in-depth review of consequences of test use, more specifically literature related to the theoretical and empirical washback studies. The above strands of literature help identify research gaps the present study attempts to fill, provide a theoretical and empirical background for the framework the study draws on, and narrow down the research purpose to specific focuses.

Chapter 4 describes a local assessment use argument within the context of the CET-4. It starts by framing the research questions in relation to the corresponding claims in the AUA template, and then proceeds to justify why claims pertaining to interpretations, decisions and consequences are the major focuses and how the generic AUA template is adapted to the locally specific CET-4 context. This chapter concludes with the articulation of the detailed warrants and rebuttals in accordance with the specific sub-research questions.

Chapter 5 reports on findings generated from a preliminary study. A statistical comparative analysis was conducted with pre- and post-2006 CET-4 to examine to what extent the current CET-4 could serve as an improved measure of students' overall English proficiency. The needs analysis explored whether the CET-4 could better reflect test takers' target language use domain and better meet students' demands for higher English proficiency. In addition, both student and teacher questionnaires were piloted and further modified.

Chapter 6 gives an overall description of the research design. It first justifies why a mixed-method approach is appropriate for collection of evidence. It then provides a sound rationale for various methods adopted, including qualitative analyses of test content, interviews, and quantitative analyses of questionnaires and test data. Finally, a detailed account of the participants, instruments, data collection and analysis procedures is presented in this chapter.

Chapter 7 reports on backing for claim of interpretation. Statistical analyses of students' CET-4 scores are presented to examine the internal structure of the test. The interpretations of scores are discussed with reference to the CECR, the 2006 CET-4 Syllabus, and publications by the NCETC. Content analyses of seven test papers actually used in the past official CET-4 administrations are described from characteristics of input and characteristics of the expected response to examine content representativeness and content relevance of the revised listening and reading components in the current CET-4.

Chapter 8 presents backing for the decision claim. Decisions that are made based on CET-4 scores by the NCETC are discussed first, followed by decisions made at institutional levels by the University Academic Affairs Office. Related backing was collected from interviews, questionnaires and document analysis. Factors underlying these decisions are discussed from perspectives of values and equitability. The chapter also identifies the potential rebuttals threatening the legitimacy of the decision claim.

Chapter 9 reports on backing for the consequence claim. It examines consequences of using the CET-4 and of the decisions made on its scores, in particular the CET-4 washback on the learning and teaching practices. This chapter reports on statistical analyses of questionnaires by addressing variables like stakeholders' perceptions of the CET-4, test preparation activities, teaching and learning behaviors.

Inferential statistical analyses were also performed to explore the relationship between test takers' perceptions of the CET-4 and their test performances. Moreover, stakeholders' interviews are reported to complement quantitative findings from questionnaires.

Chapter 10 synthesizes backing from multiple sources to revisit the AUA for the CET-4. It summarizes major findings for the research questions and discusses the implications of the study. Moreover, it identifies the limitations of the study and proposes suggestions for future research.

CHAPTER 2

THE EVOLUTION OF THE CET-4

2.1 Introduction

This chapter traces the evolution of the CET-4 from its inception in 1987 through to its largest reform in 2006. Documentation of the National College English Teaching Syllabus (NCETS) and the CET-4 Syllabus, analyses of test designers' publications, and reviews of scholarly articles related to the CET-4 help delineate the planning, development and administration of the test and better contextualize the research background. In this chapter, section 2.2 presents a general introduction to the CET-4 including its test purposes, population and test organization. Section 2.3 reviews the test development before 2006. Section 2.4 details revisions made in the CET-4 since 2006, covering its reform background, the current test contents and formats. Section 2.5 reviews validation and washback studies on the CET-4.

2.2 General introduction to the CET-4

College English Test (CET) is a nationwide standardized test, administered by the NCETC under the leadership of the Higher Education Department, the Ministry of Education (Jin & Yang, 2006). The CET is a test battery, consisting of two levels, the CET Band Four (CET-4), and the CET Band Six (CET-6). Both levels have their corresponding CET Spoken English Test (CET-SET). The present study is targeted at the CET-4 written test, and the following part will introduce its test purpose, population, and test organization.

2.2.1 Test purpose

Bachman (1990) believes that the primary concern in both test development and score interpretations is the intended test purpose. He stresses that the fundamental purpose of a language assessment is to collect information to help us make decisions about test takers' language ability (Bachman & Palmer, 2010). According to Hughes (1989, p.8), the major types of test purposes can be summarized as measuring students' language proficiency, examining learning outcome, diagnosing students' strengths and weaknesses, and making placement decisions.

A clear understanding of the test purpose is a foundation to a better vision of the test construct and washback. Therefore, the first step to review the CET-4 is to examine its intended test purpose. In accordance with the NCETS (issued in 1985/1986, revised in 1999), and the CECR issued in 2004, the CET-4 has been used as an assessment tool to examine whether college students have met the requirements of the compulsory College English course of Band 4 as specified in the NCETS (Jin, 2010; Zhang, 2008). As discussed in Chapter 1, the intended purpose of the CET-4 is to provide an objective evaluation of a student's overall English proficiency and positively impact EFL teaching at the tertiary level in China (e.g., Jin 2005, 2006, 2008; Jin & Yang 2006; Yang & Weir 1998).

2.2.2 Test population

The CET-4 is targeted at non-English major students at the tertiary level in China. University undergraduates are expected to develop their English proficiency in listening, reading, writing, and speaking after two years' English compulsory course. Students usually take the CET-4 during Year 2, but some key universities permit their students to take the test at the end of Year 1. If students fail to pass the cut scores of the CET-4, they are not eligible to take the CET-6. However, students are permitted to retake the CET-4 without restrictions on when to take it and how many times they can take it until their graduation.

According to statistics revealed by the NCETC (Jin, 2006, 2010; Yang & Weir, 1998), the CET population increased from 100,000 students in 1987 to 4.6 million in 2003. The annual CET test population soared to 10.5 million in 2005, 12 million in 2006 and 12.5 million in 2007. By 2006, over seven million students had obtained CET-4 certificates, and two million students had obtained CET-6 certificates. The CET has become the largest EFL test in the world (Jin, 2006, 2010). The dramatically increasing number of candidature in the past two decades also embodies the large-scale and high-stakes nature of the CET-4.

2.2.3 Test organization

National College English Testing Committee

The NCETC is in charge of the design, development and administration of the CET. The Committee was originated from the College English Test Design Group, which was formed in 1986 by the State Education Commission (now the Ministry of Education), soon after the promulgation of the National College English Teaching Syllabus. This Group, consisting of 12 professional language teaching and testing experts from different universities across China, designed and developed the CET-4 — the first standardized English language test for tertiary institutions in China. In 1994, the Design Group was formally recognized and established as the National College English Testing Committee. The Committee currently comprises 25 professors from 23 universities in China. The Committee members meet on a regular basis for item writing, score equating and reporting, test revision, policy-making and so forth (Jin, 2010).

The NCETC has been committed to improving the test quality and promoting College English teaching and learning. In recent years, the Committee, following the current trend in the field of language assessment, has attached more significance to policies relating to test use and test impact. In addition to test development, the NCETC, as an academic organization, has been actively participating in various academic exchanges. In 2000, the NCETC for the first time introduced the CET to the world at the Language Testing Research

Colloquium in Vancouver. In 2002, The NCETC hosted the First International Conference on English Language Testing in China, with the conference theme on the relationship between language teaching and testing and power of the test. In 2005, the NCETC introduced revisions made in the CET-4 at the "Big Test" Symposium of the International Association of Applied Linguistics in Wisconsin (Jin, 2010). In addition to presenting at various international conferences, the NCETC has engaged in regular academic exchanges with key language testing organizations established in Hong Kong, Taiwan, Japan and Korea. The NCETC has grown into a professional testing organization and played a leading role at the forefront of English language testing in Asia.

Operational structure of the NCETC

Hierarchical with the Higher Education Department of the Ministry of Education at the top, the overall operational structure of the NCETC is a coordinated group of various departments and organizations (Jin, 2010). The following part will introduce the operational structure of the CET prior to and after the 2006 reform, including organizations subordinated to the NCETC at different levels and their major responsibilities, along with changes after 2006.

The NCETC has a professional team of item writers, essay markers, and oral examiners to ensure test quality and reliability. For test administration, the CET Administration Office affiliated to the NCETC is responsible for editing test papers, printing and distributing test materials, as well as calculating, equating, reporting and releasing scores. Before 2006, three test centers were established under the Administration Office in Tsinghua University, Shanghai Jiao Tong University and Wuhan University, correspondingly responsible for the administrative work in North and Northwest China, East and South China, and the Southwest, Northwest and Central China regions (Jin, 2010; Yang & Weir, 1998). Subordinated to Higher Education Department (HED), the head of HED of the Provincial or Municipal Education Commission was appointed as the CET chief supervisor, and the head of the University Academic Affairs Office served as CET supervisor at the institutional level (Jin, 2010).

The operational structure underwent some adjustments in the 2006 CET-4 reform. The NCETC, mainly as an academic organization, is engaged in test design, quality control and score reporting, to facilitate the test operation. The National Education Examinations Authority and its subordinated Provincial or Municipal Education Examination Authority are established to execute administrative authority, including test administration and coordination of marking centers. With this hierarchical structure, the CET is expected to be administered smoothly and successfully at various levels across China (Jin, 2010).

2.3 Test development before 2006

According to Jin (2008), the major development of the CET-4 can be classified into three stages: stage 1 (1987-1996), stage 2 (1997-2005), and stage 3 (2006 onwards). Chronologically, this section will follow Jin's classification to report the major development in each stage. Since the CET-4 is designed in accordance with the NCETS and is implemented to promote College English teaching and learning, the evolution of the CET-4 should be closely related to the historical development of the NCETS and the College English teaching (see Table 2.1).

Table 2.1 Historical development of NCETS, CET-4 and College English teaching

NCETS	**College English teaching**	**CET-4**
Promulgation of NCETS (1985,1986)	Growth stage	Stage 1: Launch of the CET-4 (1987)
Revising NCETS (1999)	Improvement stage	Stage 2: Minor revisions of the CET-4 (1997) Launch of the CET-4 SET (1999)
Promulgation of the Requirement (2004, 2007)	Reforming and further development stage	Stage 3: Revising the CET-4 Syllabus Launch of the revised CET-4 (2006)

Hence, a further move is taken to situate the review of the CET-4 development in the introduction of the NCETS, the description of the features of College English education and the major trends of language assessment field in each stage.

Stage 1 (1985-1996)

During the late 1970s and the early 1980s, with China's implementation of the reform and opening-up policy and the restoration of its national college and university entrance examination, there had been an increasing demand for graduates' English proficiency. The State Education Commission (now the MoE) set up a professional team to conduct a two-year large-scale questionnaire survey to better identify students' real English levels and the social and professional requirements for graduates' English proficiency. The survey revealed that only one third of graduates met the required reading speed of 17 words per minute (Weir, Yang & Jin, 2000). The result indicated that students' reading ability required an urgent improvement. Therefore, based on the needs analysis, two National College English Teaching Syllabuses for students of science and engineering, and students of arts and humanities were promulgated in 1985 and 1986 respectively by the State Education Commission. The two syllabuses shared the same teaching objectives and requirements. The stated objectives of College English teaching were to develop students' higher ability to read in English and use English as a medium to access information related to their major areas of study, intermediate level of listening and translation abilities, and basic level of writing and speaking abilities (State Education Commission, 1985, 1986). The Syllabuses improved the required reading speed for all the students to 50 wpm for careful reading and 80 wpm for quick reading. Considering a shortage of qualified College English teachers and lower proficiency of students, the Syllabuses, which still emphasized reading ability, were regarded as a compromise between teaching requirements and teacher qualifications (Liu & Dai, 2004). Despite their focus on students' linguistic competence, the two Syllabuses met the pressing social demands at that time for improving students' access to scientific and technical literature through reading in English.

To promote the implementation of the NCETS, the CET-4 was designed and launched nationwide in 1987. A review of trends in language teaching and testing fields before the 1980s helps better understand rationales underlying the initial design of the CET-4. Before the 1980s, language ability was regarded as consisting of skills of listening, speaking, reading and writing, and components such as grammar and pronunciation. Test design focused on testing isolated 'discrete points' of language (Carroll, 1968; Lado, 1961). Later Oller's (1979) hypothesis that language proficiency comprised a single unitary trait had been prevalent for a period of time. Meanwhile, reliability was widely viewed as a prerequisite for validity. This overarching concern with the marking reliability in the field of psychological measurement made multiple-choice questions (MCQ) the dominant task type adopted by large-scale standardized tests around the world. At the earlier stage, the CET-4 followed the psychometric structuralist approach in its test design. Under the influence of the componential view of language construct, the CET-4 was designed with different sections assessing the major language skills (Jin &Yang, 2006): Listening (20%), Reading (40%), Vocabulary and Structure (15%), Cloze (10%), and Essay Writing (15%). About 85% of items were multiple choice questions and the Essay Writing (15%) was the only part testing productive skills. In line with the emphasis of the NCETS on developing students' reading ability, the Reading component occupied the largest weight (40%).

In 1990 a minor reform was made to separate the Essay Writing part from the whole paper as CET-4 Test Paper II. The purpose of this modification was to prevent test takers' overreliance on getting scores from the multiple choice questions or abandonment of the writing part. Paper I with the objective items would be collected upon the time limit and the last 30 minutes was left to students only for the writing part.

Stage 2 (1997-2005)

The 1985/1986 Syllabuses had been guiding College English teaching until 1999 and its contribution to the restoration of EFL education at tertiary level could never be denied. Since the implementation of the NCETS and the CET-4, the College English education has made rapid growth. In order to push College English teaching into a new phase and meet the challenges posed by the 21st century, the MoE issued the revised version of the NCETS in 1999. This revised Syllabus adjusted the teaching objectives to cultivate students' strong reading ability and the intermediate abilities in listening, speaking, writing and translation to enable students to communicate effectively in English (Ministry of Education, 1999). It also set a specific plan for students' four-year English study. The fundamental stage of English study was divided into four bands covering the first two years. The advanced stage (the last two years) offered students optional courses on special fields of English study.

During this period, language teaching and testing fields witnessed new trends as well. The unitary trait hypothesis of language proficiency was criticized. Instead, language ability has been viewed as multi componential and dynamic and the communicative competence model has gained acknowledgement (Bachman, 2000). A shift from psychometric-structuralist testing (to test the learner's competence through objective discrete-point items) to psycholinguistic-sociolinguistic testing (to test the learner's performance by communicative tasks) emerged (Liu & Dai, 2004). Corresponding to these changes and new trends, the NCETC took a series of measures to improve the CET-4 including more scientific and accurate score calculations, launch of the Spoken English Test, and adoption of more constructed response items.

In 1997, the Average Graded Score was introduced. Its calculation is based on the grades of the passers' scores as well as the cumulative passing rate rather than on the simple percentage of students whose scores are beyond the minimum cut-off line. The adoption of the Average Graded Score ensures a more accurate and scientific evaluation of the overall performance of a target population at different levels and prevent the pursuit of a high passing rate by teachers and a narrow pass by students (Jin, 2006, 2008).

In 1998, the Writing part set a minimum score requirement to call for more attention to cultivating students' EFL writing ability. Certain points will be deducted from the total score if a test taker's writing score is below the minimum requirement. A zero in writing will fail the test taker regardless of his performances in other components.

The launch of the CET-4 SET in 1999 is regarded as a milestone in its development. The SET is expected to serve as an appropriate tool to evaluate students' oral English and to exert a much-needed positive washback on classroom teaching of oral communication in English (Jin, 2000, 2006). The test is conducted in the form of face-to-face interview with interactive tasks such as answering the interlocutor's questions and group discussions. The CET-SET is administered biannually (in May and November). Only test takers whose scores in pre-2006 CET-4 were beyond 80 could apply for SET, and currently beyond 550 in post-2006 CET-4. The cut-off line is to limit the number of test takers to a manageable size. A large number of CET-SET test centers have been established in 28 provinces and municipalities in China, and the annual number of candidates has reached over 90,000 (Jin, 2006).

Since 1997 a variety of constructed response items have been introduced to remedy the inadequate assessment of students' productive skills (Jin, 2008). Compound dictation was an alternative for three listening passages. Short answer questions (SAQ) and Translation from English to Chinese served as the alternatives for Cloze. These new task types were expected to reduce the proportion of MCQ and improve test validity. Nevertheless, they were in fact seldom adopted. The five traditional components, namely Listening, Reading, Vocabulary and Structure, Cloze and Writing, were still dominant task types. Table 2.2 illustrates contents and formats of the CET-4 before 2006.

Table 2.2 Contents and formats of the pre-2006 CET-4 (adapted from Jin, 2008, p.7)

Component	Test content	Test format	items	Scores	Time
Listening	10 short conversations	MCQ	20	20%	20 min
	3 Passages *or* 1 Compound dictation	MCQ / gap-filling			
Reading	4 passages for in-depth comprehension	MCQ	20	40%	35 min
Vocabulary & structure	30 sentences for vocabulary and structure	MCQ	30	15%	20 min
Cloze/SAQ/ Translation	1 cloze passage *or*	MCQ	20	10%	15 min
	5 SAQ *or* 5 sentences for English to Chinese translation	SAQ/ Translation	5		
Writing	A short essay of 100 words	writing	1	15%	30 min
Total	75%-85% MCQs 15%-25% Writing and constructed response items		91	100%	120 min

To sum up test development before 2006, just as Yang and Jin noted (2000), different views on language and language ability decide the differences in test content and test method. When defining the construct of the CET-4, test designers took into account every school of thought in the field and attempted to incorporate their merits. Discrete-point items on separate skills, items on integrative skills, and items on Communicative Language Ability all occupied certain proportions. Vocabulary and Structures were to measure the candidate's linguistic competence, using discrete point separate skills items; Cloze was to measure the candidate's performance ability, using integrative skills items. The Listening and Reading Comprehension parts were to measure the receptive abilities of language use. The Writing part was to measure Communicative Language Ability in written form (Yang & Jin, 2000).

2.4 Test development after 2006 (Stage 3: 2006 onwards)

Since 2006, the CET development has ushered into its third stage. The NCETC launched its largest CET reform in 2006. This section will describe the reform background, the revised contents and formats in the current CET-4.

2.4.1 Reform background

Through the development in the past decades, the CET has achieved higher validity and reliability, met the international standards of educational assessment, and promoted the research and practices of language testing and assessment in China (Jin 2006, Jin & Yang, 2006). The CET-4 has successfully promoted the implementation of the NCETS and hence, the improvement of College English teaching and learning (Jin & Yang, 2006; Wu, 2005; Yang & Weir, 1998). It has strongly motivated both students and teachers to put more focus on their English teaching and learning. Teaching resources and facilities such as books, audio-visual materials and language labs are more readily available and accessible to teachers and students (Jin, 2010). Statistics from different administrations of the test indicate that university students have made rapid progress in their English proficiency. Meanwhile, the data provide feedback for educational policy-makers and administration authorities at various levels to make appropriate decisions with respect to English language teaching and learning (Jin, 2010; Jin & Yang, 2006; Yang & Weir, 1998). A CET certificate has become a nationally recognized credential for employment of college and university graduates (Jin, 2008; Jin & Yang, 2006; Yang &Weir, 1998). Wu Qidi, Vice-Minister of MoE, remarked in the press conference on CET-4 reform that with 17 years' steady development, the CET-4 had met social needs, won social recognition, produced beneficial effects on society, and contributed significantly to the continual improvement of the quality of college English teaching in China (Wu, 2005).

However, in spite of the painstaking efforts the NCETC has made to improve the CET-4 quality and the tremendous contribution of the CET-4 to College English teaching and learning, dissatisfaction with and criticism about the test has emerged at the same time. Many critics accused the CET-4 of a clumsy copy of the TOEFL, suffering from the inherent defects of multiple-choice items (Liu & Dai, 2004). The large proportion of MCQ was severally criticized in the 1990s for encouraging guessing and easy cheating (Hughes, 2003). The CET-4's overreliance on MCQ and accurate machine scoring of these objective items ensured high reliability but failed to measure test takers' real communicative proficiency. In terms of the test components, much criticism centered on lack of authenticity in listening and reading materials, heavy grammatical check of cloze items, restrictions of a given title and a Chinese outline on students' writing content and structure, etc (Cai, 2002, 2006).

Since the 1990s, language testers have shifted their attention from test reliability and validity to the beneficial washabck of an assessment on language teaching and learning. Hence, more criticisms stemmed from the unintended washabck of the CET-4. The high-stakes nature of the CET-4 has induced phenomenon of "teaching and learning to the test". Normal teaching was replaced by test preparation activities at the close time of the CET-4 administration. Once students got a certificate, they even gave up their College English learning. Teachers' freedom to teach creatively was constrained and great pressure was imposed on them with the passing rate. Students complained that the energy-consuming test preparations resulted in little progress in listening and speaking abilities but in good mastery of test-wiseness strategies. Students were frequently subjected to great anxiety and stress induced by the unintended uses of CET-4 scores. Students' CET-4 performances can determine whether they can obtain a Bachelor's degree, a good job, or even a residence permit in some major cities (Cai, 2005, 2006; Gu, 2003; Jin, 2008).

Due to the strong criticisms on its test design and negative washback, an urgent call for the CET-4 reform was universally perceived to promote effective teaching and learning, and to follow the new trends in language assessment field. In addition, with China's entry into WTO and success in bidding for the 2008 Olympic Games, there was also a surging need for graduates with a good command of English, especially higher communicative proficiency in English. To cater such a pressing social demand and further improve College English teaching quality, the Higher Education Department of the MoE has conducted a series of CET reforms as part of the higher education reform package since 2004. The purpose of this unprecedented reform is to provide a comprehensive assessment to measure students' communicative language ability, promote College English teaching and learning, and meet the challenges posed by social and economic development (Jin, 2006)

2.4.2 New teaching syllabus

The CET-4 is a criterion-related norm-referenced test. It is criterion-related in that its assessment on students' English proficiency is in accordance with the uniformed national teaching syllabus. Its test format and content are based on the CET-4 syllabus. Therefore, revisions of the two syllabuses are the first move of this largest reform.

The CECR stipulates: "The objective of College English is to develop students' ability to use English in an all-round way, especially in listening and speaking, so that in their future work and social interactions they will be able to exchange information effectively through both spoken and written channels so as to meet the needs of China's social development and international exchanges" (Ministry of Education, 2004, p.5). Table 2.3 highlights the major features of the CECR by juxtaposing the teaching objectives and competence development of the three teaching syllabuses. It can be seen from the comparison that cultivating students' listening and speaking abilities has been emphasized as the current focus of College English teaching.

In addition, the requirements for College English teaching are classified at basic, intermediate and higher levels and the basic requirements are set as a must for all non-English majors to achieve before graduation. Cultivation of listening and speaking abilities should always be the top priority in developing students'

English language competence at the three levels. According to the CECR, course designing should allocate adequate teaching hours and credits in cultivating students' competence in listening and speaking. Advanced information technology such as computer-based and web-based English teaching should be encouraged and promoted. Student-centered teaching model is advocated to cultivate students' language use ability and autonomous learning ability. With regard to evaluation, the CECR proposes various kinds of formative assessments such as students' self-assessment, peer-assessment, and assessment conducted by teachers and school administrators, along with the traditional summative assessment.

Table 2.3 Comparison of the three teaching syllabuses

NCETS	**Teaching objectives**	**Competence development**
The 1985/1986 NCETS	to enable students to use English as a medium to access information related to their major areas of study	higher level of reading ability, intermediate level of listening and translation abilities, and basic level of writing and speaking abilities
The 1999 NCETS	to enable students to communicate in English	higher level of reading ability, intermediate level of listening, speaking, writing and translation abilities
The 2004 CECR	to enable students to use English in an all-round way, to exchange information effectively in their future work and social interactions	ability to use English in an all-round way, especially in listening and speaking

2.4.3 Revised CET-4 syllabus

In February 2005, the Ministry of Education held a news conference on the revision of the CET, at which Wu Qidi, Vice-Minister of MoE, addressed the background, necessity, and blueprint of this reform. In March 2005, the Ministry of Education issued the trial version of the Reform Blueprint of the CET (*Quanguo Daxue Yingyu Siliuji Kaoshi Gaige Fang'an).*

In November 2006, echoing the CECR and the 2005 Reform Blueprint, the NCETC released the College English Test Band 4 Syllabus (2006 revised version) and initiated the largest reform on the CET-4. The syllabus briefly addresses aspects of test specifications, administrations and scoring procedures. It consists of a statement of the test purpose, descriptions of all the task types and the assessed language abilities, lists of language skills tested by each part, and a prototype test.

In December 2006, the CET-4 with the reformed test contents and formats was launched nationwide. Corresponding to the NCETS's advocacy of using computer-based and web-based multimedia teaching model, the NCETC commenced the project of the web-based CET in 2007 and the internet-based CET-4 was administered to students from 50 universities on a trial basis in December 2008.

2.4.4 The post-2006 CET-4

Bachman (1990) proposes five features to describe a given language test: the intended purposes or uses, content, frame of references, scoring procedure and testing methods. Since some of them have been discussed in previous sections, the following part will briefly reexamine the first four features in section 2.4.4.1, and then introduce the test method in detail in section 2.4.4.2 to delineate a comprehensive description of the post-2006 CET-4.

2.4.4.1 General features of the post-2006 CET-4

There has been a long debate on whether the CET-4 is a proficiency test or an achievement test. Test designers and administrators claim that since its inception the CET-4 has been designed as an achievement test to measure whether students' English proficiency has met requirements set by the NCETS (Yang & Weir, 1998, p.59). Test content should be based on the uniform CET-4 Syllabus designed in accordance with the uniform teaching syllabus.

Based on the CECR, the CET-4 is developed with the purpose to better measure students' overall English proficiency and to maximize its positive washback on College English teaching and learning. In spite of the long existing debate on whether the CET-4 is a proficiency test or an achievement test CET designers and administrators still refer to the revised 2006 version as an achievement test.

With regard to frame of reference, the CET-4 is a criterion-related norm-referenced test. Norm-referenced tests discriminate between test takers and rank them. Test results are interpreted with reference to the performance of a given group, or norm. The norm group is typically a large group of individuals who are similar to the individuals for whom the test is designed. Criterion-referenced tests determine how an individual performs with respect to a criterion level of ability or domain of content (Allison, 1999; Bachman, 1990). The CET-4 norm group consists of about 10,000 college students from the top six universities in China. The reported score not only indicates whether a candidate has met the requirements of the teaching syllabus for Band 4 students, but also what the percentile position a candidate occupies in the norm group (Jin, 2006).

The scoring procedure of the CET-4 consists of machine scoring of traditional MCQ and rater marking of subjective items. For the current CET-4, the Essay Writing with the newly added constructed response items have amounted to 30% to 45% of total weight. Scores from every administration of the CET-4 undergo equating and normalization before released to test takers. The score reporting system of the CET-4 was reformed by the NCETC in June 2005. Test takers used to be awarded a certificate indicating a pass (60 points or above) or a distinction (85 points or above) on the hundred point score scale which had a mean of 72 and a standard deviation of 12. Since 2005, the scores have been reported on a 710 score scale with a mean of 500 and a standard deviation of 70 (Jin, 2006, 2008). There is no pass or fail. Test takers with scores beyond 220 will be issued a Score

Report Form (SFR) with total and profile scores: total score (710 points, 100%), Listening Comprehension (249 points, 35%), Reading Comprehension (249 point, 35%), Cloze or Error Correction (70 points, 10%), Writing and Translation (142 points, 20%). The NCETC made the decision to replace the CET certificate with a new Score Report Form in hope of encouraging appropriate uses of test results and reducing social pressure imposed on the CET-4 (Jin, 2008).

2.4.4.2 Specific features of test methods

Bachman believes that test performance can be affected by the characteristics of the methods used to elicit test performance, and he further develops a framework for delineating the specific features of test method, which covers five categories: setting, assessment rubric, input, expected response, and the relationship between input and response (Bachman, 1990; Bachman & Palmer, 1996, 2010).

Students take the CET-4 in classrooms of their home institutes so that their performance will not be affected by unfamiliar environment. An earphone to receive radio signal is required for Listening component. The whole test is administered from 9:10am to 11:20a.m, lasting for 130 minutes.

In terms of the test rubric, Table 2.4 summarizes test content, format, weight, and time allocation of each component. The CET-4 consists of four components: Listening, Reading, Cloze, Essay writing and Translation. Compared with the pre-2006 CET-4 (see Table 2.2), it is in Listening and Reading components that the most significant modifications are made. As stated earlier, new task types of Long conversations and Compound dictation are included in the post-2006 CET-4, and the Listening weight is increased from 20% to 35% of the total weight. This reform is in accordance with the teaching objective of cultivating students' overall English proficiency especially their listening ability. For Reading Comprehension, a 10% of Skimming and scanning and a 5% of Banked cloze substitute for two of the in-depth reading passages in the pre-2006 CET-4. The weight of Reading component is decreased from 40% to 35%. The traditional component of

Vocabulary and Structure was cancelled. Components of Cloze and Essay writing remain unchanged but a 5% of Chinese to English translation is included into the Writing part. One sequence change is that students are required to finish essay writing and fast reading printed on separate paper at the beginning of the test administration. In the following part, each test component will be elaborated respectively.

Table 2.4 Contents and formats of the post-2006 CET-4 (adapted from Jin, 2008, p.7)

Component	Test Content		Test Format	Item		Weight		Time
Listening	Conversations	8 Short conversions	MCQ	8	15	15%	35%	35 min
		2 Long conversions	MCQ	7				
	3 Short Passages		MCQ	10	20	20%		
	1 Compound dictation		Gap-filling	10				
Reading	2 short passages for in-depth reading		MCQ	10	20	20%	35%	40min
	1 passage for vocabulary knowledge		Banked cloze	10		5%		
	1 long passage for Skimming and Scanning		T&F/MCQ, Sentence completion	10	10	10%		
Cloze	Cloze		MCQ	20		10%	10%	15min
Writing& Translation	Writing		Essay writing	1		15%	20%	30min
	Translation		Chinese to English	5		5%		5 min
Total	55%-70% MCQs 30%-45% Writing and constructed response items			91		100%		125 min

Input refers to the material contained in the task, which test takers or language users are expected to process in some way and to which they are expected to respond. It consists of format of input and language of input. The expected response consists of the linguistic or non-linguistic behavior the assessment task is attempting to elicit by the instructions, rubric and the input provided (Bachman& Palmer, 2010). Given the purpose of the study and the exhaustive and meticulous taxonomies of Bachman and Palmer's framework of task characteristics, the description of each test component will be built on major features of input and the expected response facets.

Listening

The Listening component is designed to measure students' ability to understand and interpret the spoken English. All the materials are in the social and academic domains and delivered in standard British or American English at the speed of 130-150 wpm. Items aim to measure test takers' a range of abilities from inferring the meaning or intentions of speakers, identifying main ideas or specific details, to understanding idiomatic expressions and grammatical structures. Thus, candidates are expected to acquire an intermediate level of listening competence and understand the listening materials on familiar topics. It takes about 40 minutes to administer 35 items in three sections.

Section A consists of eight one-turn short conversations, and two long conversations. The long conversation is a new test method. It is a dialogue with multiple turns and has a length of 200 words. This task type shares more resemblance to real-life situations and is expected to improve the authenticity of listening part.

Section B presents three short passages, and each passage is around 200-250 words long followed by three or four questions. These two sections contain twenty-five four-option MCQ items. The input is broadcast only once and test takers can hear questions after listening to the material, which imposes a heavy reliance on their memory, even though note-taking is permitted.

Section C delivers a Compound dictation for three times. It requires test takers to supply seven exact missing words and three missing sentences based on their understanding. This new task not only examines student's productive ability in the form of constructed response items, but also assesses their listening ability at lexical and discourse levels.

Reading

The Reading component is designed to measure students' ability to understand and interpret the written English. The total weight is decreased from 40% to 35% in the current CET-4. It consists of two parts, fast reading (skimming and scanning, 10%) and careful reading (reading in depth, 25%). The input materials are mainly of an EAP (English for academic purpose) nature, varied in topics of humanities, social science, natural science, and in genres of argumentation, exposition and narration.

Skimming and scanning is a newly added test method. Skimming is intended to examine test takers' ability to get the main idea. Scanning is to examine their ability to locate specific information by cues such as a figure, a capitalized word or words at the beginning or the end of a paragraph. Test takers are required to read a 1000-word-long passage at the speed of 100 wpm and finish seven true or false questions, which are replaced by MCQ in recent years, and three gap-filling or sentence completion items within 15 minutes.

Reading in depth contains two sections. Section A is Banked cloze, a new task type to replace the component of Vocabulary and Structure. Test takers are required to pick out appropriate words from a bank containing 15 word choices to fill in 10 blanks in a 200-250 word-long passage. Banked cloze is intended to assess contextualized language use instead of context–free knowledge of language. It requires not only knowledge of vocabulary and structure but also the skill of inferring contextual meaning of vocabulary at the discourse level (Jin, 2008).

Section B requires test takers to read two long passages (300-350 words per passage) at a speed of 70 wpm and finish 10 MCQ items. It is intended to measure students' reading competence at various levels including understanding of gist and details, making inferences and figuring out vocabularies from context.

Cloze

Cloze has been an established test method since the inception of the CET-4. It examines students' integrative ability to understand and use language at lexical, semantic and contextual levels. Test takers are required to read a passage of 220 to 250 words and fill in 20 blanks from MCQ items within 15 minutes. It takes up 10% of the total weight.

Writing & Translation

The Writing part measures test takers' productive ability to express their ideas in written English. It requires test takers to produce an essay of no less than 120 words within 30 minutes on a single prompt of a given topic, a set situation, a picture or a diagram. Test takers are instructed to write each paragraph based on a given topic sentence in Chinese. It occupies 15 % of the total weight.

The content domain is mainly related to campus and academic settings, or social issues which students are familiar with. The writing genres are expected to cover argumentation, exposition, narration, letters or other forms of practical writing. Argumentation is the frequently adopted genre for CET-4 writing which requires test takers to take a position and defend it, or analyze a problem and propose the possible solutions.

Translation part requires test takers to complete five sentences (15-30 words per sentence,) by translating the phrases or fragment sentences from Chinese to English within 5 minutes. It measures students' grammatical and lexical knowledge, as well as their ability to use idiomatic expressions appropriately and accurately. It constitutes 5% of weight. Scores of translation and writing are added up together and presented in the Score Report Form.

CET-4 on screen marking was put into trial in 2003 and has been implemented in all test centers since 2006. The essay marking adopts a holistic approach in light of the large size of test population. One rater rates each essay and gives reward scores based on the global scoring and his holistic impression. The well-established uniform scoring criterion ensures the intra-marker, inter-marker, and inter-center consistency. In addition, the final writing score is subject to computer adjustment to filter out inconsistencies resulting from marker subjectivity. The NCETC released that CET marking reliability reached to 0.87, indicating that the CET writing scores can be interpreted as an indicator of candidates' writing ability (Jin, 2006).

Speaking

The CET-4 SET is optional for students. Only students whose scores are beyond the cutoff line can take it. The SET is composed of three parts and has 20 minutes of administration. The test is conducted in the form of interviews and group discussions. Both authorized examiners evaluate and score the candidate's performance independently based on three criteria: accuracy and range, size and discourse management, and flexibility and appropriateness.

2.5 Previous studies on the CET-4

This section will review previous studies on validity and washback of the CET-4. First, the only large-scale and influential validation study conducted by the NCETC will be reported in terms of its aim, methodology, findings and conclusions. Second, major empirical studies on washback of the CET-4 will be summarized. The purpose of this section is to reveal what has been done and what methodology or frameworks have been adopted and to explore what can be further explored to niche research gaps in validation and washback studies.

2.5.1 The 1998 CET-4 validation study

An empirical and large-scale CET validation study that must be discussed is the project jointly conducted by the NCETC and the British Council in the late 1990s. This project lasted for three years and investigated the CET from multi-facets of validity including content validity, concurrent validity, predictive validity, and face validity.

Aim

This validation project was intended to fulfill several research purposes: 1) to examine whether the CET-4 and the CET-6 can be regarded as reliable, scientific and accurate measures of university students' language proficiency; 2) to develop new task types to improve the test format and content; 3) to improve the positive washback effects on College English teaching and learning at the tertiary level; 4) to examine the reliability of the statistical analysis for test data; 5) to promote the theoretical research on language testing and large-scale standardized tests (Yang & Weir, 1998, p.53).

Methodology

During the three years' research, test specifications were documented and prototype tests were constructed. Statistical analyses on the internal structure of the test including correlation and factor analyses at item and total test levels were conducted to examine construct validity. For criterion-related validity, large comparative tests were administrated. The CET-4 and the CET-6 were compared for predicative validity, and concurrent validity was examined by a comparative study between the CET-6 and Japanese Education Ministry and Society for testing English Proficiency (STEP). For content validity, CET test papers from 1987 to 1995 were analysed in the aspect of genres, topics and readability, etc. The required test taking skills were examined in accordance with the test specifications. An introspective study involving 40 students was conducted with reading components of the CET-4 and the CET-6 prototype tests. Marking reliability was also one of the foci of the study. Questionnaires and interviews were adopted for complementation of the quantitative analyses.

Findings & conclusions

Yang and Weir (1998) revealed that the internal correlations among different components in their validation study fell within the range of 0.3 to 0.7. They explained that higher correlation between two components in a test indicated their measuring the same language ability, so it was unnecessary to keep both. By contrast, lower correlation meant the two components measured different abilities, so both failed to contribute to an overall ability that the test was intended to measure. Therefore, they interpreted their correlations, without being too high or low, as appropriate and acceptable in language testing field and claimed that CET subtests measured different but interrelated skills (Yang & Weir, 1998, p.60). Cohen and Holliday (1982) suggest the yardsticks for correlations: 0.19 and below is very low; 0.2 to 0.39 is low; 0.40 to 0.69 is modest; 0.70 to 0.89 is high; and 0.90 to 1 is very high. Based on these criteria, Yang and Weir concluded that the correlation coefficients revealed by the CET validation study can be viewed as acceptable. Yang and Weir (1998, p.60) also conducted exploratory factor analysis, which revealed that different subtests all contributed to one factor measured by CET, which they interpreted as general linguistic competence. Pre-testing on the CET established norms and a comprehensive check of the coverage of test operations assured the appropriate and consistent item difficulty and discrimination power of CET. Evidence from introspective study revealed that there was a relationship between students' test performance and the application of reading strategies expected by test developers. The investigation of CET marker reliability indicated that a qualified team of essay markers and a system of quality control had been established. Questionnaire surveys and interviews indicated both teachers' and students' satisfaction with the overall CET.

According to the NCETC, this validation project has shown that the CET-4 served as a valid and reliable measure of test takers' general linguistic competence, effectively promoted the implementation of teaching syllabus and objectively reflected students' English proficiency at tertiary level (Yang & Weir, 1998; Weir, Yang & Jin, 2000).

2.5.2 The CET-4 washback studies

In addition to the three-year Sino-British collaborative validation project, large-scale empirical studies concerning validity of the CET-4 by researchers outside of the NCETC were rarely found. One of the possible reasons might be the difficult access to real test data from the NCETC, which hindered more in-depth and objective investigation into the item difficulty, discrimination and other latent variables in the CET-4 by employing advanced statistical analyses such as IRT or SEM approaches. Since the early 1990s both theoretical and empirical washback studies abroad have developed boomingly (e.g., Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Andrews, 1994; Bailey, 1996; Hughes, 1989; Messick, 1996; Wall, 1996; Watanable, 1996), while there was a dearth of washback studies in China. With language testers' increasing interests in washback phenomenon in the early 21st century, both the NCETC and the individual researches attempted to investigate washback of the CET-4 on College English teaching and learning from various perspectives. Among them a few studies are worth discussion given their research scope and depth (see Table 2.5).

Table 2.5 Washback studies on the CET-4 in China

Researcher	Instruments	Subjects	Major findings
Jin (2000)	Questionnaires	358 students 28 oral raters	• It was necessary to launch the CET-SET and its positive effects on CE teaching and learning had been observed. • Teachers put more focus on cultivating students' communicative language ability. • Students were motivated to participate in oral activities.
Huang (2002)	Questionnaires Interviews Classroom observations	120 teachers 600 students	• The CET-4 exerted more positive washback than its negative effect. • The intensity of CET-4 washback differed from ordinary universities to key universities, from Year 1 students to Year 2 students. • The CET-4 washback was a result of various factors, of which test methods and test policies played an important role.
Gu (2004, 2007)	Questionnaires Interviews Classroom observations Document analysis	4500 stakeholders including administrators, teachers and students	• The CET-4 has made great contribution to CE teaching and learning, and its positive washback outweighed the negative. • The CET-4 washback was more salient on teaching content, pace and attitudes than on teaching methods. • The CET-4 washback varied with different university types, grade cohorts, and teacher factors.

Jin (2000) conducted a washback study on the CET-SET, involving 358 SET takers and 28 SET raters from universities in Shanghai, Beijing and Nanjing. Questionnaires as the major instrument were administered to respondents to explore their attitudes on necessity, test design, methods of the SET, and its actual and potential influences. Raters were asked to further evaluate the rating scales. Findings indicated that the SET was strongly welcome and its positive washabck had been observed. Teachers put more focus on cultivating students' communicative language ability and students were motivated to participate in oral activities as well. The majority of respondents showed positive attitudes on the test method and administration.

Huang's (2002) washback study was targeted at the CET-4 written test, involving 120 teachers and 600 students from nine universities in Chongqing and two universities in Chengdu. In addition to questionnaires, he also employed interviews and classroom observations to complement each other. Findings indicated that positive washback of the CET-4 outweighed its negative influences, and the intensity of CET-4 washback differed from ordinary universities to key universities, from freshmen to sophomores. The study concluded that CET-4 washback was a result of various factors, of which test methods and test policies played an important role.

Gu (2004, 2007) carried out a large-scale empirical study on the CET-4, in which 4500 stakeholders were recruited from 391 colleges and universities across China, including test administrators, CET markers, college English teachers and students. Triangulating abundant evidence from questionnaires, interviews, classroom observations, and document analysis, the study revealed that most stakeholders acknowledged the great contributions of the CET-4 to College English teaching and learning, and believed its positive washback outweighed the negative. Their dissatisfaction mainly stemmed from the overuse of MCQ, use of coaching materials in class and suspension of normal textbook teaching for test preparation. The CET washback was more salient on teaching content, teaching pace and teachers’ attitudes than on teaching methods. In addition, the CET-4 washback varied with different university types, grade cohort, and teacher factors.

Since 2007, small-scale studies targeting at the current CET-4 emerged. Some investigated the washback phenomenon from teachers’ perspectives, focusing on teaching content, teaching methods, and teacher factors (Li, 2008; Peng, 2009), while others explored students’ perceptions (Huang, 2009; Hou & Wang, 2008). However, since the current CET-4 was launched only in 2006, there has been a dearth of large-scale and profound studies on it. Hence, further research on its validity and washback is urgently needed.

2.5.3 Critique of previous CET-4 studies

Even though the above major studies either pioneered China’s washabck research or pushed its development, a close examination on them also revealed certain limitations: 1) Questionnaire analysis only involved basic descriptive statistics such as percentage or means to reveal respondents’ opinions. Few studies endeavored to link respondents’ test performances to their perceptions of the CET-4 or learning behaviors by employing regression analysis. 2) Researchers tended to investigate either teaching or learning perspectives and teachers were usually the major studied subjects. Few considered using the independent t-test to compare teachers’ and students’ attitudes towards similar questions so as to find out any possible differences and explore underlying reasons. 3) Washback mechanism is far more complicated than collection of the attitudinal opinions and observation of classroom teaching and learning behaviors. The above studies failed to establish an explicit and operational framework to guide their studies except that Gu (2005, 2007) extended Hughes’s PPP model (participants, processes, and products) to PPPP model (participants, perceptions, processes, and products). 4) The NCETC followed an older fashion of validity in conducting their CET validation studies. Above all, the above-mentioned washback studies did not investigate washback by incorporating it into validation as Messick advocated in his validation framework. They separated washback from validity in their investigation into and discussion of such a large-scale and high-stakes test. Hence, further research is needed to keep pace with the latest development and trends in the field language testing. The research gap revealed by the above-mentioned limitations in previous studies is just what the present study aims to fill in.

2.6 Summary

This chapter has reviewed the evolutionary stages of the CET-4 corresponding to revisions made by the NCETS at each stage. The review covered the general features of test purposes and methods, test organization and operation, in particular the reform background and the revised elements in the current CET-4. In addition, some major and influential studies on the CET-4 were summarized in the latter part of this chapter. It can be seen that the two decades between 1985 and 2005 witnessed an enormous improvement in EFL education at the tertiary level in China, and the NCETC has always been taking pains to improve the test qualities so as to better serve College English teaching. On the other hand, the CET-4 is also much debated for its high-stakes nature. In response to higher demands for students' overall English proficiency, the emerging negative effects of the test, and the new trends in language testing development, the CET-4 underwent an unprecedented reform in 2006 in order to make the test more communicatively oriented and maximize its positive washback effect.

To sum up, the purpose of describing the evolution of the CET-4 is to familiarize readers with the research context, statement of the problem, and motivations of the present study. Moreover, reviewing domestic studies on the CET-4 helped disclose some research gaps within China's EFL assessment context. The present study is intended to draw on an AUA, the latest framework, to address validity issues of the CET-4 and its washback by a series of inference from score interpretations to score-based decisions, and to test consequences. The next chapter will extend the literature review to a broader scope, taking a historical approach to delineate the evolution of validity and washback mechanism to lay a solid rationale for the AUA framework and draw further relevance to the present study.

CHAPTER 3

LITERATURE REVIEW

3.1 Introduction

The purpose of this chapter is to provide conceptual, theoretical and methodological underpinnings for the present study. The first strand of literature takes a historical approach to delineate the evolving conception of validity and validation approaches developed. The second strand provides a basic review on consequences of test use, or more specifically on washback. By reviewing the existing literature, this chapter identifies gaps and research focus, as well as provides insights for the research design.

3.2 Evolving conception of validity

The theoretical conception of validity has gradually evolved over the years (Anastasi, 1986; Angoff, 1988). Given its changing meaning and broad scope, this section, structured chronologically, is intended to highlight some important aspects to this evolution so that the contemporary concept of validity will be better appreciated against the historical perspective.

3.2.1 Historical view of validity as a componential concept

Criterion-based validity model

Early discussions of validity can be traced back to the beginning of the 20th century. Validity was defined as the correlation of test scores with "some other objective measure of that which the test is used to measure" (Bingham, 1937, p.214), or "in a very general sense, a test is valid for anything with which it correlates" (Guilford, 1946, p.429). Cureton (1951) in the first edition of *Educational Measurement* defined validity in terms of the correlation between the actual test scores and the "true" criterion scores (p.623). It can be seen that validity was evaluated in terms of how well the test scores estimated or predicated the criterion scores. It was mainly viewed as a concept of validity coefficient, dominantly defined in the form of a correlation and in its predictive sense. Once a criterion is specified, the criterion model provides a simple, elegant, and effective approach to validation. After data on some sample of individuals are collected, a validity coefficient can be computed in a straight way (Cronbach & Glester, 1965; Cureton, 1951). By the early 1950s, the criterion-based model had been well developed and widely applied. Just as Cronbach (1971, p. 443) said "The theory of predication was very nearly the whole of validity until about 1950", criterion validity was valued as the gold standard for validity (Angoff, 1988; Cronbach, 1971; Moss, 1992; Shepard, 1993).

However, the major limitation of the criterion-based model lies in the fact that it may be difficult to implement a criterion that is clearly better than the test itself or even to conceptualize a satisfactory criterion (Cronbach, 1971, 1980; Guion, 1998; Lord & Novick, 1968). The fundamental problem that plagues criterion-related validity studies is whether there is an existing valid criterion or how the criterion can be validated. Evidence for the validity of the criterion itself requires its correlation with other tests, or other indicators of ability, which leads to infinite circularity in comparing the test to criterion A, and criterion A to criterion B, etc, resulting in an endless spiral of concurrent relatedness (Bachman, 1990; Kane, 2001). Thus, the criterion model does not provide a good basis for validating the criterion.

Content-based validity model

In response to shortcomings of criterion-based model, namely, unavailability of reliable and meaningful criteria, content-based validity model was developed to establish the plausibility of criterion measure. The rationale is to employ a criterion measure involving some desired performance and interpret the scores in terms of this kind of performance. Content validity offers a way to validate the criterion by establishing a rational link between the procedures used to generate the criterion scores and the proposed interpretation or use of scores (Cureton, 1951;

Ebel, 1961). Where a sample of some type of performance is used to draw conclusions about level of skill in that kind of performance, a good case for validity of the proposed interpretation can be made on rational grounds (Cronbach, 1971; Cureton, 1951; Ebel, 1961; Kane, Crook, & Cohen, 1999). The content model interprets test scores based on a sample of performances in some area of activity as an estimate of overall level of skill in that activity. This approach tends to work especially well for tests of specific skills, and it has most frequently been applied to measures of academic achievement (Flockton & Crooks, 2002).

Later a significant modification was made on the characterization of content validity in the 1974 *Standards* published by American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on measurement in Education (NCME), changing from content domain to behavior domain. To be specific, content validity in the 1966 *Standards* (APA, AERA, &NCEM, 1966) was described as how well "the test samples represent the situations or subject matter about which conclusions are to be drawn", whereas the 1974 edition described it as how well "the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in the desired domain" (p.28). The behavior a test is intended to elicit calls for some construct to define. Thus, the content validity and construct validity become integrated. Such a change also has methodological implications. It is not evidentially sufficient to say whether a test is valid only grounded on the expert judgment that a test represents a certain domain of situations or subject matter. Investigation into the behaviors of examinees should be a vital source of evidence as well.

Nevertheless, this view of building validity of the test on a review of the test content by subject-matter experts is also under a number of criticisms. First, content-related evidence "tends to be highly subjective and has a strong confirmatory bias" (Kane, 2001, p.320) when relying on judgments about the relevance and representativeness of test tasks, especially judgments produced by test developers. Second, demonstrating either content relevance or content coverage is difficult since the domain specification is more than a simple list. Moreover, content-related evidence only supports interpretations that are limited to the domain specified, and this limited interpretation tends to be unidirectional (Bachman, 1990; Messick, 1980). The primary limitation of content validity is that it focuses on tests, rather than test scores. The content of a given test does not vary across groups of examinees, but the performances of these individuals may vary considerably, and the interpretations of test scores will vary accordingly (Bachman, 1990; Messick, 1975).

To sum up, content-related evidence provides support for the domain relevance and representativeness of the test instrument. Demonstrating a test is relevant to and covers a given area of content or ability is therefore a necessary part of validation (Bachman, 1990; Messick, 1989). However, content validity model has a limited role in validation in that it does not provide direct evidence for the inferences to be made from test scores (Messick, 1989). It is even problematic when it is used to argue for the validity of claims about cognitive processes or other theoretical constructs, since demonstrating the contents of a test accurately represent a given domain of ability does not take into consideration how individuals actually perform on the test (Cronbach, 1971). Thus, content validity provides necessary but insufficient source of evidence in test validity. Further examinations into test takers' internal or cognitive processes are tended to be employed as a supplement.

Construct-based validity model

Historically, the notion of construct validity emerged in the early 1950s. During this period different researchers employed a confusing array of names to report their findings on validity, ranging from Guilford's (1946) factorial and practical types, Cronbach's (1949) logical and empirical types, to Anastasi's (1954) face validity, content validity, factorial validity, and empirical validity. Angoff's (1988) historical review of conceptions of test validity describes no fewer than 16 types since the 1930s. To clarify the chaotic state and confusing conceptions on validity, and in response to limitations of criterion-based and content-based validity, the

Technical recommendations for Psychological Tests and Diagnostic Techniques was published (APA, 1954), in which Meehl and Cronbach introduced the notion and terminology of construct validity. They further developed this concept in their classic article in 1955, adopting the hypothetico-deductive (HD) model of theories to describe the need for construct validation and provide the conceptual framework for its investigation:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypothesis, which are a means of confirming or disconfirming the claim (Cronbach & Meehl, 1955, p.284).

Cronbach and Meehl explicated concept of construct validity in detail in their subsequent articles (Cronbach & Meehl, 1955, Cronbach, 1971, 1980a, 1980b). Three aspects of construct-based model gradually emerged as general principles of validation, applicable to all proposed interpretations (Kane, 2001). First, Cronbach and Meehl (1955) stated explicitly that validation of an interpretation always involved an extended analysis, including the development of theory, the development of measurement procedures, the development of specific hypotheses based on the theory, and the testing of these hypotheses against observations. Thus, the construct validity model highlighted the inadequacies of most validation efforts based on a single validity coefficient or simply on expert opinion (Cronbach, 1971). Second, Cronbach and Meehl (1955) increased awareness of the need to specify the proposed interpretation before evaluating its validity, saying that "the network defining the construct, and the derivation leading to the predicted observation, must be reasonably explicit so that validating evidence may be properly interpreted" (p.300). Compared with the criterion model, the construct model shifted the emphasis from the validation of the test to the development and validation of a proposed interpretation. It is not the test or the test score that is validated, but a proposed interpretation of the score (Cronbach, 1971). Third, construct validity's focus on theory testing suggested the need to challenge proposed interpretations or to evaluate the competing interpretations (Cronbach, 1971; Cronbach & Meehl, 1955), which was largely ignored in the criterion and content models. As Kane (2001) summarized, the construct-validity model has pushed the validation theory forward by developing three methodological principles: the need for extended analysis in validation, the need for an explicit statement of the proposed interpretation, and the need to consider alternate interpretations, in the context of validating theoretical constructs (APA, 1954; Cronbach & Meehl, 1955).

However, construct validity model also has some inherent demerits. First, before the 1980s it was essentially viewed as an addition to the criterion models, just as what Cronbach and Meehl (1955) described that "construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned and must use indirect measures" (p.283). They suggested construct validity is a pervasive concern, but did not present it as a general organizing framework for validity. The 1966 *Standards* (APA, AERA, & NCEM, 1966) distinguished construct validity particularly from criterion validity, but still presented construct model as an alternative to the criterion and content model rather than an overriding concern. The 1974 *Standards* (APA, AERA, &NCEM, 1974) continued this trend in tying construct validity to theoretical constructs. Second, the HD model residing in the logic structure of theories and justifications as interpreted axiomatic system, results in the limited application of construct validity to areas where there is less solid theory. Likewise, the difference distinguished between the strong program and the weak program of construct validity (Cronbach, 1988; Cronbach & Meehl, 1955) leads to much confusion. The strong program calls for well-established theories to undergird it, otherwise it would be of limited utility. The weak program is inclined to pull all the evidence under a unified concept, which may result in opportunistic choice of evidence rather than the most relevant evidence if no explicit guidance is provided. Hence, as Kane (2001) comments, construct validity has not provided a unifying influence on an operational level, and the criteria for evaluating validity evidence were still in doubt.

Section summary

The aforementioned three parts have introduced criterion, content, and construct validity respectively, and in fact, these three types of validity as a whole are called the Trinitarian model of validity, which had been dominant during the late 1970s and the early 1980s and even has a great influence up to now. However, the tripartite categorization of validity has had some adverse side effects on testing practice. Essentially, it represents "a crude and oversimplified grouping of many data-gathering procedures that contribute to an understanding of what a test measures" (Anastasi, 1986. p.2). Hence, the Trinitarian model was challenged in the 1980s and rejected with the publication of the 1985 *Standards*. I would like to present the tendency of how the three types of validity came to be seen as a unitary concept of validity by summarizing the evolution of the *Standards* cited earlier.

The 1954 *Standards* initially listed four types validity, namely, concurrent, predictive, content, and construct validity. Later the 1966 *Standards* amalgamated predictive validity and concurrent validity into criterion-related validity. The tripartite categorization of validity first has been regarded as three distinct types of validity. Each model was employed as needed in test validation. The criterion model was used to validate selection and placement decisions. The content model was used to justify the validity of various achievement tests. Construct validation was to be used for more theory-based, explanatory interpretations (Kane, 2001). However, in most cases, more than one model was called for to provide validity evidence. Therefore, the 1966 *Standards* noted, "The three aspects of validity are only conceptually independent, and a complete study of a test would normally involve information about all types of validity" (AERA, APA, & NCME, 1966, p.14). The third edition of the *Standards* (AERA, APA, & NCME, 1974) retained the three types of validity but characterized them as "interdependent kinds of inferential interpretation" and as three essential aspects or components of validity that are operationally and logically interrelated. The 1974 *Standards* stressed that "Questions of validity are questions of what may properly be inferred from a test score: validity refers to the appropriateness of inferences from test scores or other forms of assessment" (AERA, APA, & NCME, 1974, p.25). In the 1985 *Standards* the categories were abandoned and the unitary interpretation became explicit (Fulcher & Davidson, 2007).

Even though Cronbach and Meehl failed to view construct validity as an overarching concept under which the criterion and content models should be subsumed, great debts should be owed to them for pushing validity to shift from its predicative sense to a more profound and manifold scope. As Bachman (1990) remarks, construct validity, since Meehl and Cronbach (1955) firstly addressed it in their seminal article, has come to be recognized by the measurement profession as central to the appropriate interpretation of test scores, and provides the basis for the view of validity as a unitary concept. In next section, I will describe how construct validity evolved from an alternative to the criterion and content model to an overarching concept.

3.2.2 Construct validity as a basis for a unified validity concept

Early in the late 1950s Loevinger (1957) had initially suggested "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636). By the late 1970s, two opposing trends were evident in the development of validity theory. One tended to maintain a clear specification of the kinds of evidence needed to validate particular interpretations and uses of test scores. Meanwhile, there was a perceived need to develop a unified conception of validity (Kane, 2006). Most theorists have reconsidered the status of the multiple definitions of validity. Rather than enumerating various types of validity as appear above, the concept of construct validity has been widely agreed upon as the single, fundamental principle that subsumes various other aspects of validation (Cumming, 1996, p.5). Seeking to resolve this tension, the fourth edition of the *Standards* (AERA, APA, & NCME, 1985) defined validity as follows:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are

> many ways of accumulating evidence to support any particular inference. Validity, however, is a unified concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (p.9)

The 1985 *Standards* rejected the traditional Trinitarian model and called the three categories as "types of evidence" rather than "types of validity". It viewed validity as a unified concept while recognizing that different kinds of evidence were relevant to different kinds of interpretations. This terminological change acknowledged a single unified view of validity with construct validity as central. Content-related and criterion-related evidence are viewed as methods for investigating construct-related evidence (Brennan, 2006, p.2; Chapelle, 1999, p.256). Thus, by the 1980s the construct validity model became widely accepted as a general approach to validation and as the basis for a unitary framework of validity (Anastasi, 1986; Guion, 1977; Messick, 1975, 1980, 1988, 1989).

Section summary

To sum up, the centrality of construct validity should be attributed to Cronbach and Meehl, who set out a theory of construct validity that impacts upon the way we think about and do test today. Their contribution to construct validity cannot be overestimated. It was their seminal paper of 1955 that introduced the term to educational and psychological testing. It contains all the ideas that have led to our expanded understanding of validity, including Messick's unified concept of validity and even validity as an argument to support claims (Fulcher & Davidson, 2007, p. 181).

So far I have presented a historical view on how validity has developed from the familiar classifications into content, predictive, concurrent and construct validity to the tripartite division, from viewing construct validity as a central to view it as a basis for a unitary concept. These concepts survive in current testing standards and guidelines, with some important shifts in emphasis. They are presented above in their classic version to provide a benchmark against which to appraise the import of subsequent changes (Messick, 1989).

3.2.3 Messick's unified validity framework

The 1980s witnessed concept syntheses of validity. The unitary concept of validity in the 1985 *Standards* concurred with what Messick endorsed in his articles appearing between the later 1970s, and the early 1980s. Validity as a unified concept had not been universally recognized until Messick proposed his framework supported by a series of argumentation particularly in his perhaps most important article on validity in 1989.

Messick (1988) criticized the 1985 *Standards* for accepting the idea that different validation efforts might involve different types of evidence, which may encourage reliance on very limited and perhaps opportunistically chosen evidence for validity. In addition, the traditional three categories of validity evidence are fragmented and incomplete in that it fails to take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use. He further asserted that the heart of the unified view of validity is that appropriateness, meaningfulness and usefulness of score-based inferences are inseparable and that the unifying force is empirically grounded construct interpretation (Messick, 1988). Therefore, in the third edition of *Educational Measurement* Messick (1989) defines validity as follows:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment…. Validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and us…. Thus, the key issues of test validity are the interpretability, relevance, and utility of scores, the import or value implications of scores as basis for action, and the functional worth of scores in terms of social consequences of their use. (p.13)

To meet these requirements Messick (1989, p.20) produced a unified validity framework by distinguishing two interconnected facets of the unitary validity concept. As Figure 3.1 displays, the left column demonstrates one facet that is the source of justification for testing. Empirical evidence for construct validation can be distinguished as evidential basis and consequential basis. The first row displays the other facet that is the function or outcome of testing. Arguments for construct validation can be distinguished based on analyses of test interpretation and test use. If the facet for justification (either an evidential basis for meaning implications or a consequential basis for value implications of scores) is crossed with the facet of for function or outcome (either test interpretation or test use), a four-way progressive validity matrix is presented, highlighting both score meaning and value implications in both test interpretation and test use (Fulcher & Davidson, 2007; Messick, 1989, 1995).

	Test interpretation	Test use
Evidential Basis	Construct validity	Construct validity + Relevance / Utility
Consequential Basis	Construct validity + Value implication	Construct validity + Relevance / Utility + Value implication + Social consequences

Figure 3.1 Messick’s progressive validity matrix (1989, p.20)

The evidential basis for test interpretation is construct validity, which requires evidence from any source to support or weaken the intended score meaning. Additional evidence for construct validity can also be derived from consideration of test use that is buttressed by evidence for the relevance of the test to the specific applied purpose and for the utility of the test in the applied setting. Let’s proceed to the second row, to justify a particular interpretation of the test score, not only all the evidence supporting or weakening the intended score meaning should be collected, but the value implications of the interpretation should be considered with the specific reference to the context for which the test is intended to use. Hence, the consequential basis of test interpretation calls for evidence concerning the theory and philosophy underlying the test, or evidence as an indicator of the individual’s ability. Such an interpretation should also take into account the value implications of various labels about what is important or valued in performance on the test. The consequential basis of test use is the appraisal of both potential and actual social consequences of the applied testing, impacts on social systems and values, including unintended, negative effects (Bachman, 1990; Fulcher & Davidson, 2007; Messick, 1989, 1995).

This fourfold classification of facets of validity manifests the overall influence of construct validity and its importance in each facet. The progressive matrix interpreted validity as a unitary but multifaceted concept. He incorporated a social dimension of assessment such as value implications and social consequences overtly into validation framework, greatly enhancing our understanding of the construct validity. The way Messick defines validity has become the accepted paradigm in psychological, educational and language testing. This can be seen from the latest 1999 *Standards* (AERA, APA, & NCME), which follows Messick (1989) closely and defines validity as:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated. (p. 9)

In addition to his emphasis on evaluating value implications and social consequences of score meaning and use in the construct validity framework, Messick (1995) pointed out two major threats to construct validity. One source of invalidity is construct underrepresentation occurring when assessment content is not reflective of relevant knowledge. The other is construct irrelevant variance referring to extraneous, uncontrolled variables that affect assessment outcomes. These two categorized sources inform how to further enhance test validity and reliability and cast illuminating implications to consequential aspects of construct validity.

The major question raised about Messick's progressive matrix pertains to the relationship between evidential basis and consequential basis of validity, which is left as an incomplete synthesis of evidence (Markus, 1998; McNamara & Roever, 2006). Bachman (2005) later argues that Messick's framework does not provide practical guidance on how to conduct validation research. In addition, it should be pointed out that at that time the most controversial aspect of Messick's unitary concept pertained to the role of consequences of test use in validation, which from the critics' perceptive should not be encompassed in validity (Mehrens, 2005; Popham, 1987). The consequences of test use or misuse deserve investigation but "involve social and moral analyses beyond the scope of test validation" (Wiley, 1991).

Section Summary

In retrospect, since the 1950s validity has moved from the early types of validity to the present unitary concept. Development, modification, and refinement have been made on the evolving property of validity, which either fundamentally changed or illuminatingly enhanced our understanding of the validity theory. Messick's unified validity framework sets out the prelude to modern conceptions of validity. The 1999 *Standards* endorses Messick's theory and reiterates that validity is a unitary concept. From the above elaboration, several aspects of the current view of validity deserve to be highlighted.

First, validity is an overall evaluation of the plausibility of a proposed interpretation or uses of test scores. What needs to be validated is the interpretation of inferences from test scores and decisions to be made on scores rather than the test or test scores per se (Bachman, 1990; Cronbach, 1971, 1988; Cureton, 1951; Cronbach & Meehl, 1955; Messick, 1975, 1989, 1995).

Second, as noted earlier, current definition of construct validity should not only encompass the proposed interpretations, but talso ake into account possible competing interpretations so as to better evaluate the adequacy and appropriateness of the interpretations (Cronbach, 1971; Messick, 1975, 1989).

Third, validity is viewed as a unitary concept with construct validity as the central, subsuming content, criterion-related evidence along with evidence from consequences of test uses. Hence, justifying the validity of a test is no long solely limited to test designers or test researchers. Test users and decision-makers should also be accountable to justify their uses and decisions they make.

Finally, validity is viewed as an argument concerning evaluation of test interpretation and test use. Since the 1980s a trend has emerged to structure validation research in terms of validity argument (Cronbach, 1980, 1988; House, 1980). Cronbach took the form of evaluative argument to discuss the validation of score interpretations and uses. He suggested *validity argument* is to provide an overall evaluation of the intended interpretations and uses of test scores by generating a coherent analysis of all of the evidence for and against the proposed interpretation/use, and to the extent possible, the evidence relevant to plausible alternate interpretations and decision procedures (Cronbach, 1988). Later the 1999 *Standards* (AERA, APA, & NCME), concurring with Cronbach's discussion, embodies the view of validation as argument by interpreting it as "Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (p.9).

In short, in the context of evolving notions of validity, Messick's validity theory has a powerful impact on and far-reaching implications for language testing research. Illuminated by his framework, quite a number of researchers on validity have elaborated and expanded on his approach. Some of the leading figures and their argument-based approach to validation will be discussed respectively in the following section.

3.2.4 Argument-based approaches to validation

Viewing validity as argument has been suggested in the 1980s, as discussed earlier, by researchers like Cronbach (1988), House (1980), and Messick (1989). However, only in recent years building an integrated and coherent evaluative or validity argument for interpretations of score meaning and uses has gradually received full attention. Given that the unified construct validity fails to provide guidance or criteria in its implementation, an argument-based approach to validation is proposed and developed, to offset the drawbacks, by specifying the intended interpretation and use as the framework for validation. In this section, three leading figures will be introduced and their argument-based approaches to validation will be elaborated. First, I will gloss Toulmin's (2003) description of informational or practical arguments, from which current approaches toward developing interpretative and validity arguments derive their structure and logic derivations.

3.2.4.1 Toulmin's argument approach

Toulmin (2003, p.8) states that "A sound argument, a well-grounded or firmly backed claim, is one which will stand up to criticism, one for which a case can be presented coming up to the standard required if it is to deserve a favorable verdict". Figure 3.2 provides the layout of an argument as outlined by Toulmin (2003, p.97).

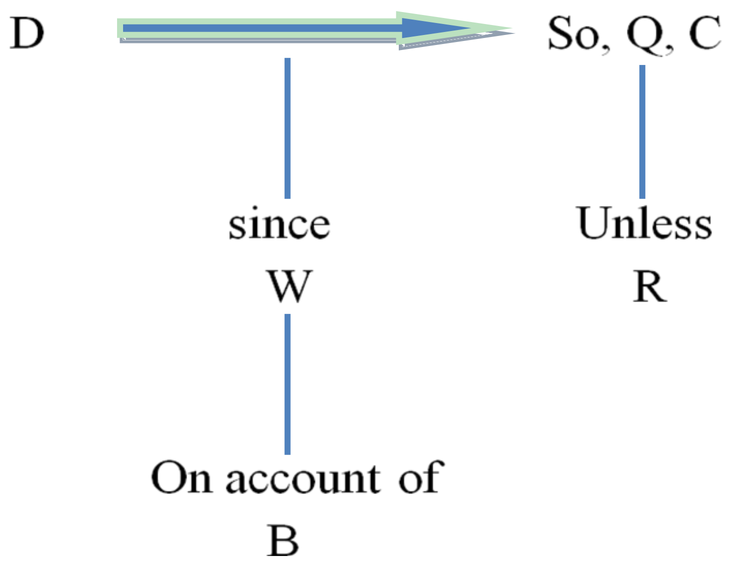

Figure 3.2 Toulmin's argument structure (Toulmin, 2003, p.97)

A basic argument structure consists of five elements: claim (C), qualifier (Q), data (D), warrant (W), rebuttal (R), and backing (B). A claim is a conclusion whose merits we are seeking to establish. The data consist of the facts we appeal to as a foundation for the claim. A warrant is a general statement that provides legitimacy of a particular step in the argument. A modal qualifier indicates the force of the warrant. The backing consists of other assurances, without which the warrants themselves would possess neither authority nor currency. A rebuttal consists of exceptional conditions which might be capable of defeating or rebutting the warranted conclusions. The rebuttal data consist of evidence that may support, weaken, or reject the alternative explanation (Toulmin, 2003). Toulmin's schema for the structure of arguments provides useful terms and representations for analyzing assessment arguments (Mislevy et al, 2003).

Now let us interpret what each element refers to when applying Toulmin's argument model in formulating an assessment validation argument. In assessment, a claim is an interpretation we intend to make for score meaning, about the knowledge, skill or ability of a test taker. Data are the responses of test takers to specific test items or tasks, or to put it simply, test performance. A warrant is the rationale or proposition usually pertinent to language and testing theories in support of the claim. The arrow between data and claim represents an inferential link between test tasks and test takers' response to the tasks, which can be justified on the basis of a warrant. Backing provides support for warrant, which can take a variety of forms including theories, prior research and both qualitative and quantitative methodologies frequently adopted in language testing. A rebuttal is the challenge to the validity of a claim, similar to an alternate or competing interpretation.

3.2.4.2 Mislevy's evidence centered design

Mislevy (1996) and colleagues (Mislevy, Steinberg, & Almond, 2002, 2003) framed Toulmin's argument structure to an approach for test designing and validation, which he called an *evidence centered design* (ECD). An ECD encompassing key elements of claims, evidence, and tasks, endeavors to establish a chain of reasoning between the claims and the evidence on which claims are based. Claims are meaningful statements of the relevant inferences we want to draw about test takers to serve an assessment's purpose. Evidence refers to observations we would need about test-takers as a basis for drawing those inferences. Tasks are activities test takers would engage to provide such evidence.

An ECD underscores the central role of evidentiary reasoning, which helps integrate construct definitions, characteristics of assessment tasks, and the psychometric models that are needed to deal with complex performance data (Bachman, 2005). The chain of reasoning entails a detailed logical analysis of the stages in test designing or validation. A preliminary stage is domain analysis, which involves developing the conceptual and organizational structure of the target domain. The second stage is domain modeling, involving potential claims about students and the aspects of proficiency they reflect, evidence about what students do or say, and tasks to elicit evidence from students. The further stage of ECD is the conceptual assessment framework (CAF), which proceeds the test design from the thinking stage to the actual operational assessment by laying out the interrelations among models as student model, evidence models, task models, assembly model, presentation model and delivery model (McNamara & Roever, 2006; Mislevy et al, 2003;). And herein I will not detail the nuts and bolts of these models since demonstrating this high-level structure suffices for a brief introduction to the structure and elements of this ECD approach as an assessment argument.

In brief, an ECD offers a framework to "transform the argument to an operational assessment by first working through the structure and then designing elements that can be assembled" (Mislevy et al., 2002, p.479). The structure of the framework makes the validity argument more explicit, while the designing objects, or models, within the framework guide the practical work of designing tests. This logical analysis of claims, evidence and tasks leads to the development of test specifications and plans for the systematic collection and analysis of test data to investigate empirical evidence in support of claims and the threats to them (McNamara, 2006). It has practical benefits in terms of guiding the development of performance tasks and scoring rubrics, and laying the foundation for collecting evidence for validity and generalization (Mislevy et al., 2002). As Mislevy et al (2003) comment, the ECD framework is useful for both analyzing existing assessments and designing new ones, but the latter should prove more immediately useful. Hence, the development of new TOEFL has drawn on an ECD to establish a validity argument as guidance in its test design process (Chapelle, Enright, & Jamieson, 2008). However, the major limitation of Mislevey's framework lies in its ignorance of the social dimension of an assessment and issues of fairness, or the issue of social values and consequences of test use that Messick concerns (Bachman, 2005;McNamara, 2003; McNamara & Roever, 2006).

3.2.4.3 Kane's interpretative argument

Kane (1992) also develops a systematic approach to drawing inferences from test scores by gathering and disseminating evidence supporting intended score interpretations, which he called an interpretative argument. Kane, Crooks and Cohen (1999) use a metaphor of bridges to demonstrate a chain of inferences in this validation process for linking observations to interpretations. Later Kane (2001, 2002, 2004) extends the linkages in his interpretative argument by addressing the role of test use in validation. Kane (2001) divides an interpretative argument into two parts: descriptive part and prescriptive part. The descriptive part links scores to statements about individuals, while the prescriptive part links these descriptive statements to the decisions that are made. Kane (2002) further delineates the interpretative argument in terms of different kinds of interpretations (descriptive and decision-based) and the kinds of assumptions (semantic and policy) on which these are based. In a subsequent article, Kane (2004) outlines an argument-based approach to validation as two steps. First, an *interpretative argument* specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the test scores, and then evaluates the coherence of argument and the plausibility of its inferences and assumptions. Second, the *validity argument* provides an evaluation of the interpretative argument and generally involves extended analysis and empirical studies. Hence, Kane's argument comprises four types of inferences from the observed performance to interpretations and further extends to decisions. Figure 3.3 illustrates the chain of inferences in this interpretative argument.

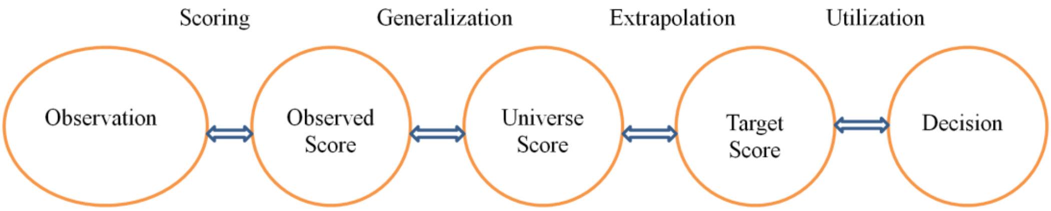

Figure 3.3 Links in an interpretative argument (Kane, 2004, p.167)

The first inference *scoring* is from an observation of test performance to a score based on assumptions about appropriateness and consistency of the scoring procedures. The second inference *generalization* is from the observed score to the universe score based on assumptions about the representativeness of the observations and generalizations of observed score obtained from test performance across tasks, judges, and occasions. The third inference *extrapolation* is from the universe score to a target score based on the claims of what a test taker knows or can do and the evidence supporting these claims. This link is akin to what the first cell in Messick's validity matrix involves, that is, the theory-defined construct. The fourth type of inference *utilization* is from target score to score-based decisions, which is in line with cells of relevance, values and consequences in Messick's validity matrix.

To date Kane's work has been routinely cited in Bachman's recent work on Assessment Use Argument (Bachman, 2002, 2003, 2004, 2005, Bachman & Palmer, 2010). Chapelle, Enright, and Jamieson (2004, 2008) have drawn on Kane's work to build a validity argument for TOEFL iBT in order to complement Mislevy's evidence centered design (McNamara, 2006). As Chapelle et al (2008, p.9) argue, compared with Mislevy's dual grounds of both observation and task characteristics, Kane (1992) suggests that multiple types of inference connect observations and conclusions. Given complexity of the TOEFL interpretative argument, distinguishing among the types of inference is critical for organizing the key processes and results into a coherent validity argument. Thus, they organize their argument in terms of the assumptions and evidence associated with six types of inference: evaluation, generalization, extrapolation, explanation, decision-making and representation (Bachman, 2005). From their empirical experience in building a validity argument for TOEFL iBT, Chapelle et al (2010, 2012) outline four advances realized in Kane's framework. First, Kane's interpretative argument provides conceptual tools to express the multifaceted meaning of test scores and downplays the need to define the construct, since the score interpretation is based on the interpretative argument rather than the construct. Second, the interpretative argument guides the validation research in a systematic manner to examine the chain of inference rather than according to a traditional list of validity evidence. Third, the results are synthesized in support of the interpretative argument, which combined constitute the validity argument. Fourth, the interpretative argument allows for the plausible alternative interpretations to challenge, weaken or refute the validity argument.

In brief, the prominent advance in Kane's work is the fourth inference moving the interpretative argument from interpretation of score meaning to actual score use. In this sense, Kane's notion of validity is in accordance with that of Messick, addressing test use and test consequences as aspects of validity. The difference lies in Kane's emphasis on conceptualizing validity as an argument. In Kane's formulation, the validity argument provides an overall evaluation of the intended interpretations and uses of test scores. The goal is to provide a coherent analysis of all of the evidence for and against proposed interpretations/uses, and to the extent possible, the evidence relevant to plausible competing interpretations (Brennan, 2006). Moreover, a distinctive emphasis in Kane's work is the need to systematically anticipate the main threats to the validity of those inferences so that appropriate validation studies may be designed to investigate those threats

(McNamara, 2006). Compared with Mislevy's work that does not engage with the issue of the consequences of test use, Kane does in principle embrace a concern for consequences. However, the question remains as to how to set about investigating test uses and consequences for which Kane failed to develop a methodology.

3.2.4.4 Bachman's Assessment Use Argument

In this part, we will first get a glimpse of Bachman's earlier work before elaborating on his latest Assessment Use Argument in hope of highlighting advances and merits of an AUA. Then, the rationale, elements and qualities of this framework will be introduced.

Bachman's work on validity has been influenced by Messick's approach to validity. In his *Fundamental Consideration in Language Testing* (1990), one of the most influential books on language testing, Bachman outlined the implications of Messick's validity as a unitary concept for language testing as well as test use and social consequences as aspects of validity. Bachman, following Messick, introduces validity as a unitary concept pertaining to test interpretations and use, emphasizing that the inferences made on the basis of test scores, and their uses are the object of validation rather than the tests themselves (Chapelle, 1999, p.257). The prominence of this book not only lies in Bachman's model of Communicative Language Ability as a theoretical framework to describe test performance, but also in his description of characteristics of test methods to elicit language performance. McNamara (2006) summarized three aspects embodied Bachman's thoroughgoing adoption of Messick's approach within the field of language assessment. First, the criterion domain is explicitly treated as a construct through analysis of the target language use domain. Second, the elaborated model of Communicative Language Ability explicates the relationship between test construct and the criterion construct. Third, test method is treated as an aspect of test content, so that test method is made to approximate the conditions of performance of target language use situation, and is thus fully accounted for. Thus, a clear relationship exists between target language use situation, test task, and test construct

(Communicative Language Ability). One effect of such clarity is that hypothesized relationships can be validated from test performance data. However, in fact, Bachman's test methods framework has been seldom implemented in test development projects given its daunting and meticulous facets.

Later Bachman and Palmer (1996) proposed an overarching notion of *test usefulness* as a more manageable validation framework, encompassing six qualities: reliability, construct validity, authenticity, interactiveness and practicality. The merit of this formulation is that it brings considerations of construct validity and impact under a unitary concept of test usefulness, while its limitation lies in lack of an explicit linkage among the six qualities or between validity and test use (Bachman, 2005). The price of manageability is a certain loss of theoretical coherence (McNamara & Roever, 2006).

Bachman (2005) indicates that the fields of language testing and educational and psychological measurement have not yet developed a comprehensive of principles and procedures for linking test scores and score-based interpretations to test use and the consequences of test use. As noted earlier, although Messick's validity matrix has called for language testers' attention to test use and consequences, he did not provide guidance on how to investigate them. Likewise, the aforementioned writers on argument-based approach to validation, mainly, Kane and Mislevy, propose procedures for establishing a chain of inferences from observation of test performance to claims to be supported by evidence, but they do not address issues of test use and consequences as well. Bachman argues that the argument-based approach to validation of score interpretations should be broadened to include an argument for test use, involving two types of evidence referred to by Messick: the relevance and usefulness of the score meaning for the intended decision. Hence, Bachman (2004a, 2005; Bachman & Palmer, 2010) proposes to use Toulmin's (2003) approach to practical reasoning as a basis for articulating an Assessment Use Argument.

An AUA is an overall logical argument for linking assessment performance to use (decisions). Bachman extends Kane's methodology for developing a validity argument to include a second, parallel validation argument for dealing with test use, so an AUA includes two parts: a utilization argument, linking an interpretation to a decision, and a validity argument, which links assessment performance to an interpretation (Bachman, 2005, p.1). An AUA also provides a conceptual framework to justify the intended uses of a particular assessment. Assessment justification is defined as a process to investigate the extent to which the intended uses of an assessment are justified. This process comprises two interrelated activities: 1) articulating specific statements in an assessment use argument (AUA) that support the links between consequences and assessment performance; 2) collecting relevant evidence, or backing, in support of the statements in the AUA. The process of justification serves two essential purposes: 1) It guides the development and use of a given language assessment and provides the basis for quality control throughout the entire process of assessment development; 2) It provides the basis for test developers and decision makers to be held accountable to those who will be affected by the use of the assessment and the decisions that are made (Bachman &Palmer, 2010. p. 95).

After getting a basic understanding of its rationale and function, I will proceed to detail the components in an AUA. Figure 3.4 presents an overview of the framework, illustrating its structure, elements, inferential links, and qualities that I will interpret in turn in the following part.

Elements of an AUA

An AUA consists of the following elements: data, claims, warrants, backing, rebuttals and rebuttal backing. **Data** consist of the information on which a claim is based, usually referring to test takers' performance on an assessment as well as the assessment per se. **Claims** (represented in rectangles located in the middle of Figure 3.4) are statements about the inferences to be made on the basis of data and the qualities of those inferences. A claim thus includes two parts: 1) an outcome of the assessment process and 2) one or more qualities of that outcome. **Warrants** (represented in ovals) are general statements that help support the qualities of claims. Warrants provide specific justification and legitimacy for the qualities that are claimed of the intended consequences, of the decisions, of the interpretations, and of the assessment records. Warrants can be supported by backing. **Backing** (represented in the vertical arrow on the right) consists of the evidence in support of warrants, which comes from a variety of sources, including documents, regulations, legal requirements, theory, prior research or experience, etc (see the large vertical arrow on the right of Figure 3.4). **Rebuttals** (represented in ovals) are statements about possible alternatives to the outcomes or to the qualities that are stated in the claims. Rebuttals challenge or reject the claimed outcomes or qualities and support alternative outcomes or alternative qualities to those stated in the claims. **Rebuttal backing** (represented in the vertical arrow on the right) consists of the evidence that we need to provide to reject or weaken the rebuttals about alternative interpretations to the stated claims (Bachman & Palmer, 2010).

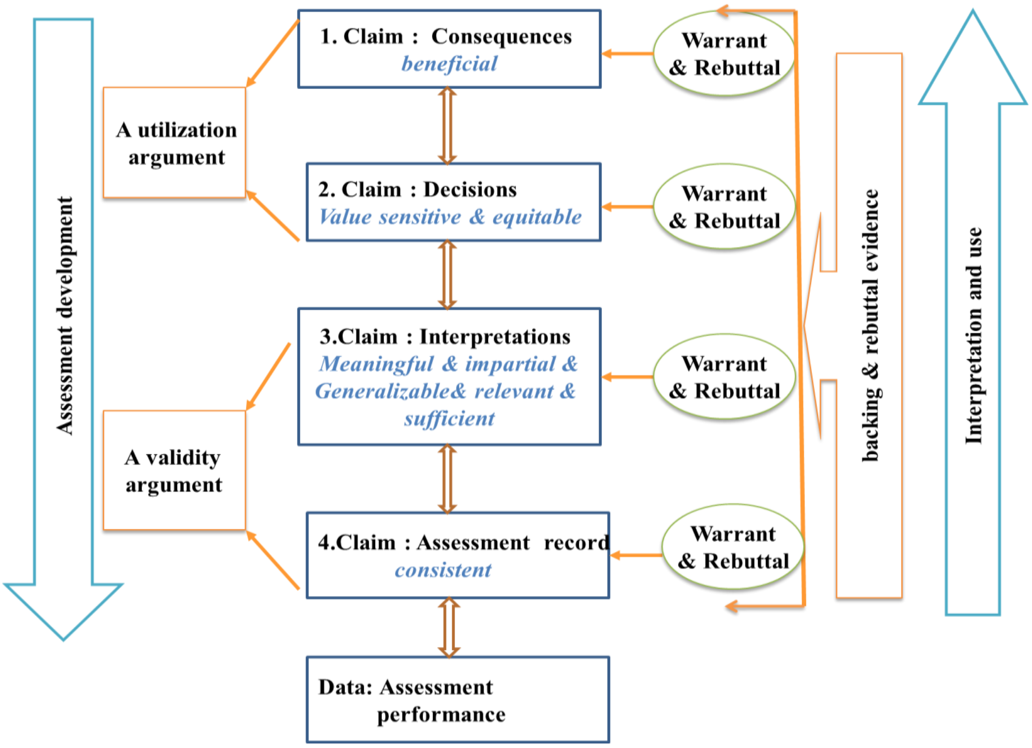

Figure 3.4 An AUA framework (adapted from Bachman & Palmer, 2010, p. 91, p104)

Inferential links in an AUA

An AUA is a conceptual framework consisting of a series of inferential links between the consequences test developers intend to bring about and the test takers' assessment performances. The downward pointing arrow on the left signifies the process of assessment development. Developing a test begins with considering what intended consequences we want to bring about. The beneficial consequences then determine decisions we would like to make on the basis of the test. Then, we delineate the ability relevant to the decisions, in other words, what aspects of ability we would like to examine, which is followed by thinking about what kind of information we need to collect that we can interpret as the indicator of an aspect of a test taker's language ability. Finally, we determine what kind of assessment tasks can help us elicit the test taker's performance we need (Bachman & Palmer, 2010). When using assessment performance to make decisions, we should follow the upward arrow on the right, making a chain of inferences from performance to assessment records, to interpretations, to decisions, and consequences.

Bachman (2005) has proposed that the basic "building block" of an AUA is a data-claim inferential link. Bachman and Palmer (2010) also suggest another way of thinking about the inferential links as a series of claims. An AUA serves as a chain of data-claim links in that the claim resulting from one inferential link becomes the data that serve as the basis for the next inference in the chain. In the lowest pair in Figure 3.4, data refers to the test taker's performance, while the claim is an assessment record, which is the score or verbal description obtained from the assessment. In the next pair, data becomes the assessment record, while the claim is the interpretation of the test taker's ability we want to assess. The interpretation in turn is data in the next pair while claim is the decision to be made. In the topmost pair decision becomes data on which the claim about the intended *consequences* of using the assessment and of the decisions that are made are based. The qualities of each claim are either supported by warrants or weakened by rebuttals in the argument. Claims on assessment record and interpretations constitute a validity argument while claims on decisions and consequences constitute a utilization argument (illustrated by two squares on the left in Figure 3.4).

Qualities of the claims

As mentioned earlier, a claim consists of one or more qualities of an outcome of the assessment process. Warrants are in fact statements in support of the qualities of the claims. The italicized words in rectangles of Figure 3.4 describe qualities proposed by Bachman and Palmer (2010) for an AUA. The introduction to these qualities begins with beneficence in the consequence claim.

Beneficence

An assessment is expected to bring about beneficial consequences to different groups of stakeholders. Kunnan (2004) interprets his principle of beneficence in such a way that a test be beneficial rather than be harmful or detrimental to society. Following Kunnan, Bachman and Palmer (2010) define beneficence as the degree to which the consequences of test uses and of decisions made on the test promote good and are not detrimental to stakeholders.

Value sensitivity & equitability

Bachman (1990) emphasizes that "tests are not developed and used in a values-free psychometric test tube; they are virtually always intended to serve the needs of an educational system or of society at large" (p.279). Therefore, decisions that are made based on an assessment should take into account the societal values like educational values of the community, values of parents, as well as relevant school regulations and legal requirements. Value sensitivity means engaging with stakeholders to understand these values.

Equitability means that decisions that are made are not biased for or against any particular group of test takers. Test takers have equal opportunity to learn the ability to be assessed, and they are classified only according to the cut scores and decision rules (Bachman & Palmer, 2010).

Meaningfulness, impartiality, generalizability, relevance & sufficiency

The meaningfulness of an interpretation mainly pertains to the construct of a test. The warrant statements involve a course syllabus, a needs analysis of language use in a target language use (TLU) domain, or a general theory of language ability, on which the construct has been defined. Warrants elaborating meaningfulness can also state that the assessment tasks engage the ability defined in the construct definition.

Warrants about impartiality are related to the fairness of assessment-based interpretations. Some warrants state that the format and content of the assessment tasks and all aspects of the administration of the assessment are free from bias that may favor or disfavor some test takers. Other warrants can state that individuals have equal access to information about the assessment itself, its content and procedures, and even have equal opportunity to prepare for the assessment (Bachman & Palmer, 2010).

Generalizability can be defined as the degree of correspondence between a given language assessment task and a TLU task in their task characteristics. Warrants addressing that the interpretation is generalizable should state the characteristics of assessment tasks, the responses of test takers to these tasks, and the interactions between them correspond to those in the TLU domain (Bachman & Palmer, 2010).

Relevance is defined as the degree to which the interpretation provides the information the decision maker needs to make a decision. This warrant underscores the importance the discussing the needs of test users or decision makers, as well as analyzing the language use domains of the TLU domain in the development of an assessment.

Sufficiency can be seen as an extension of relevance in that it addresses the question of how much relevant information is needed for the decision maker to feel comfortable so as to make correct decisions based on the interpretations derived from the assessment record.

Consistency

Consistency is the extent to which test takers' performances on different assessments of the same construct yield essentially the same assessment records. Consistency warrants state that the assessment records are consistent across different characteristics of assessment including different assessment tasks, forms of assessment, assessors, or times of assessment. Some warrants may involve administering and scoring procedures (Bachman & Palmer, 2010).

Merits and demerits of an AUA

To sum up the aforementioned discussion, an AUA consists of a series of statements (claims, warrants, rebuttals) about the outcomes (consequences, decisions, interpretations, assessment records) of a given assessment and about the qualities of these outcomes. It provides a means for defining the qualities that are associated with specific assessment outcomes as well as for understanding the relationships among these qualities. In an AUA, the qualities of consistency, meaningfulness, impartiality, generalizability, relevance, sufficiency, values sensitivity, equitability and beneficence are associated with specific claims and warrants in an AUA, and are linked conceptually by the structure of the AUA itself (Bachman & Palmer, 2010).

The merits of Bachman and Palmer's AUA lie in the following aspects. First, an AUA provides a conceptual and logical framework to guide the process of assessment development or assessment justification. Second, this framework advances the argument-based approaches to validation to include an argument for test use. Test consequences are linked to validity issues via a series of coherent inferences in an AUA. Third, rather than seek evidence according to the traditional "checklist" of validities, an AUA provides sufficient flexibility in collecting the backing evidence most relevant to validity or test use claim. Time and resources can be efficiently allocated to collect evidence only pertaining to specific warrants and rebuttals.

However, demerits can also be observed in an AUA. As a new framework, Bachman and Palmer subsume some specific terminologies (consistency, meaningfulness, equitability, and consequence) in an AUA to address conventional conceptions of reliability, validity, fairness, washback and impact. As Kane (2011, p.585) indicates, it is understandable that new terminologies are used in a new framework. However, unnecessary confusion may be caused when definition of a given term in an AUA is different from the traditional notion of the term. For instance, generalizability is conventionally related to reliability and the G-theory. Bachman and Palmer (2010, p.117) define it in a way more similar to authenticity. Another source of confusion is the role of rebuttal. In an AUA there is a potential rebuttal for every warrant. Even though Bachman and Palmer (2010) suggest that it makes more sense to argue for warrants rather than state them as implied rebuttals, it is sometimes weird that a researcher articulates a rebuttal to challenge a warrant that he just produced. It does not mean rebuttal should be ignored in an AUA. In order to avoid possible confusion, what I suggest is that a researcher should identify the most challenging rebuttals and specify who proposes them and on what ground these rebuttals are proposed.

Section summary

In this section three leading authors on argument-based approach to validation and their formulations are discussed respectively: Mislevy's evidence-centered design, Kane's interpretative argument, and Bachman and Palmer's Assessment Use Argument. Given detailed introduction in previous parts, I will not repeat the content and features of their formulations herein but discuss advantages of this approach as a concluding remark on this section. The merits of viewing validity as argument rather than considering it as a process of collecting evidence according to the traditional "checklist" are emphasized as follows. First, in the application of checklists the tendency is to look for evidence that supports the validity or test use claim, whereas in an argument approach researchers are forced to focus on disconfirming evidence (Haertel, 1996). Second, an argument draws our attention to the validity questions most relevant to a particular test and its purpose so that time and resources can be spent on collecting evidence that bears on specific warrants or rebuttals, which is especially useful when there is limited time or validation studies (Bachman, 2005, Bachman & Palmer, 2010; Haertel, 1996).

3.2.5 Backing sources

As mentioned earlier, assessment justification includes two steps. Articulating specific statements (claims, warrants and rebuttals) in an AUA constitutes the first step in the assessment justification process. The aforementioned section has informed the basic elements and features in articulating an AUA. The next step is to collect evidence or backing to support the warrants in an AUA. Backing can be collected from a wide range of sources and in a variety of ways. Some backing may be provided at the outset of or during assessment development. Some will be collected in the form of empirical evidence during the tryout and operational administration of the assessment (Bachman &Palmer, 2010). As Bachman (2004) indicates that validation can be seen as the process of articulating an argument and collecting evidence in support of a particular interpretation of test scores, the way to collect backing in this sense is the approach to collecting evidence in validation approaches. Thus, the traditional methods are still available to collect backing including questionnaires, think-aloud protocols, observations and descriptions, interviews, and statistical analyses of assessment records (Bachman & Palmer, 2010). This section will address the major evidential sources of validation.

Test validation has long been recognized as a process that requires many types of evidence, analyses, and interpretation. All data yielded by the administration of a test could serve as legitimate evidence of validity (Angoff, 1988, p.30). Bachman (1990) describes validation as “a general process that consists of the marshaling of evidence to support a given interpretation or use, a process that is based on logical, empirical and ethical considerations” (p.238). Messick (1989) outlined the following approaches, which have been viewed as indispensable and widely employed in validation studies since the 1990s:

> We can look at the content of a test in relation to the content of the domain of relevance. We can probe the ways in which individuals respond to items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal

> structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is, the test's external structure. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions — such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also the unintended side effects. (p.16)

Following Messick, Bachman (2004) later more explicitly summarizes several different approaches to validation: 1) the analysis of test content, 2) the analysis of test taking process, 3) the analysis of correlations among scores from a large number of tests, 4) the analysis of differences among comparison groups in an experimental design, and 5) the analysis of differences among non-equivalent criterion groups. The following part will address major qualitative and quantitative approaches respectively.

Qualitative approaches

Analysis of test content

Evidence in support of the claim of content representativeness and relevance can be collected by analyzing test content, providing evidence for the hypothesized match between test items or tasks and the construct that the test is intended to measure. A key issue is to address the domain specification. The accepted approach is that the experts read the test and make judgments about the cognitive knowledge and processes, skills and other attributes to be revealed by the assessment tasks through job analysis, task analysis, curriculum analysis and especially domain theory (Messick, 1989, 1995). A number of studies illustrate the application of this approach (e.g., Alderson, 1993; Bachman, Daviion, Ryan & Choi, 1995; Bachman, Kunnan, Vanniarajan, & Lynch, 1988).

Analysis of test taking process

One type of qualitative analyses attempts to document the strategies and language that learners use as they complete test tasks. The hypothesis would be that the test taker is engaging in construct-relevant processes during test taking (Chapelle, 1999). The most commonly used approach for collecting empirical evidence about the processes used in taking tests is to ask test takers to provide a verbal report of the processes they use (Bachman, 2004). Recently there has been an increasing interest in gathering verbal accounts on how individual test takers respond to assessment tasks either by think-aloud approach or retrospective verbal report. Thus, verbal protocol analysis has been viewed as an indispensible tool for collecting information in that it can provide valuable and more in-depth insight into the real process in which test takers handle assessment tasks. Studies employing this approach are conducted on tests of listening, reading and cloze tests (Anderson et al., 1991; Buck, 1991; Cohen, 1984; Nevo, 1989; Wu, 1998).

Quantitative approaches

Correlation is an extensively used approach in validation studies to support a claim that a particular test measures a particular area of knowledge or ability. It is frequently adopted to assess the internal structure of a test by correlating different test components with each other and between each subtest and the whole test (Alderson, et al., 1995). Correlational approaches to construct validation may utilize both exploratory and confirmatory modes. The approach to administer a number of different tests to the same group of test takers and then analyze the patterns of correlations among the different sets of test scores is called exploratory factor analysis (EFA). EFA is useful for investigating patterns of convergence, or commonality, among many different measures. A number of studies have used EFA to investigate the abilities that are measured by different types of language tests (e.g., Bachman et al., 1995; Carroll, 1983; Oller, 1979). The confirmatory factor analysis (CFA) begins with hypotheses about traits and how they are related to each other, and then attempts to either confirm or reject these hypotheses by examining the observed correlations (Bachman, 1990). The CFA is often used with the multitrait-multimethod correlation, an example of which is illustrated by Bachman and Palmer’ (1982) validation study related to communicative proficiency.

According to Bachman (2004), the limitation of EFA is that it looks only at convergence, the claim that scores measure the same ability, but fails to provide evidence to reject the counterclaim that other abilities or test methods may also affect test performance. Thus, the MTMM design is adopted to investigate relationships between test scores and other tests and behaviors. In this approach, each measure is considered to be a combination of trait and method, and tests are included in the design so as to combine multiple traits with multiple methods, and then evidence for validity is found if the correlations among the tests of the same construct are stronger than correlations among tests of different constructs (Bachman, 1990; Chapelle, 1999). The advantage of the MTMM lies in its examination into patterns of both convergence and discrimination among correlations, which makes this approach often used with CFA in language testing research.

One of the relatively new quantitative methodologies favored by language testers (e.g., Kunnan, 1995; Purpura, 1999, 1998; Sasaki, 1993) in recent years is structural equation modeling (SEM). Kunnan (1998b) outlined five research objectives of SEM for language testing. In brief, this approach can be viewed as encompassing several models: multiple regression, path analysis, and factor analysis, offering a mechanism to estimate and test hypothesized relationships among a set of substantively meaningful variables (Bentler, 1995; Kunnan, 1999).

Dimensionality analysis, investigates the internal structure of the test by assessing the extent to which observed dimensionality of response data is consistent with the hypothesized dimensionality of the construct. When the psychometric model is unidimensional, classical true-score reliability and item response theory methods are used to investigate the data fit (Chapelle, 1999).

The next source of evidence is drawn from results of research on differences in test performance. Hypotheses are based on a theory of the construct which includes how it should behave differently across groups of test takers, time, instruction, or test task characteristics. The study of how differences in test task characteristics influence performance is framed in terms of generalizability (Chapelle, 1999).

The final type of argument cited as pertaining to validity are those arguments based upon testing consequences. The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term. Social consequences of testing may be either positive or negative (Messick, 1995). This approach involves both qualitative and quantitative methodologies like interview, observations, and questionnaires, which will be addressed in the second strand of literature review in this chapter.

Section summary

The aforementioned types of validity evidence provide a coherent introduction to research on validation and a way of addressing the multiple and interrelated validity questions that need to be answered to justify score interpretation and use. They guide validation research which integrates evidence from these approaches into validity. In addition, whatever qualitative or quantitative approach is adopted, kinds of evidence to be collected will be guided by the specific claims, warrants and rebuttals in an assessment use argument. The evidential sources of validation, or backing sources, are addressed herein with more breadth than depth in hope of presenting an overview of the general methodologies, while backing collection approaches to be adopted in the present study will be elaborated in Chapter 6 Methodologies for the main study.

3.3 Washback

The focus of this section is on washback, which is expected to provide theoretical underpinnings for consequence claim in Bachman and Palmer’s AUA. It first presents definitions and scope of washback, involving introducing other related key terms like impact and consequence, the terminology used in Bachman and Palmer’s AUA. Then nature and mechanism of washback are discussed. It continues to examine some empirical washback studies. The remaining part addresses the role of consequence in validation. One point needs to be clarified that this section is not intended to delineate a comprehensive and in-depth review on complex concept and mechanism of washback in that our research focus is more on exploring an approach to link washback investigation to validity issues than on washback phenomenon per se. Thus, the review is expected to provide readers with a better understanding of the consequence claim and its subordinate washback warrant to be articulated in next chapter.

3.3.1 Definition and scope

Washback is generally defined as the influence of testing on teaching and learning (Alderson & Wall, 1993; Bailey, 1996). Shohamy (1992) describes washback as "the utilization of external language tests to affect and drive foreign language learning the school context" (p.513). Messick (1996), viewing washback as part of what he calls consequential validity, defines it as "the extent to which the test influences language teachers and learners to do things they would not otherwise necessarily do that promote or inhibit language learning" (p.241). Cheng (1997, 1998, 2005) defines washback as an intended direction and function of curriculum change, by means of a change of public examinations, on aspects of teaching and learning. Andrew (2004) defines washback within the educational setting as "the effects of tests on teaching and learning, the educational system, and the various stakeholders in the education process" (p.37).

A synonym of washback is backwash, which is defined by Hughes (2003) as the effect of testing on teaching and learning. Hughes explains that the two terms are interchangeable but he prefers backwash in that its meaning can be found in dictionaries. Spolsky (1994, p.55) believes that backwash carries negative connotation so "the term is better applied only to accidental side effects of examinations, and not to those effects intended when the first purpose of the examination is control of curriculum". In fact, in addition to washback and backwash, with the increasing interests in the nature and phenomenon of washback, different researchers have produced a set of terms denoting the relationship between testing and teaching and learning. *Measurement-driven instruction* has positive connotation that important tests can lead to educational improvement (Frederickson, 1984; Popham, 1987). *Curriculum alignment* refers to the concept of aligning the content of instruction to test content (Madaus, 1988; Smith, 1991; Shepard, 1993), implying negative connotation of narrowing the curriculum and teachers' training practices. *Systemic validity* means that a new or revised examination is introduced into the education system bringing about improved curricular and instructional changes (Frederiksen & Collins, 1989). Although different terms are preferred by different researchers, they all refer to different facets of the same phenomenon—the influence of testing on teaching and learning (Cheng & Andy, 2004). In field of applied linguistics, particularly language education and language testing, the term Washback is most commonly used. This study follows suit using the term washback referring to test influence within instructional setting.

Impact is another term denoting a larger scope of test influence. Bachman and Palmer (1996) define test impact at two levels: a micro level and a macro level. They conclude that "the notion of washback in language testing can be characterized in terms of impact, and includes the potential impact on test takers and their characteristics, on teaching and learning activities, and on educational systems and society" (Bachman & Palmer, 1996, p.35). Wall also made a distinction between the two concepts. Washback is "frequently used to refer to the effects of tests on teaching and learning" whereas "impact refers to any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole"(Wall, 1997, p.291). Hamp-Lyons (1997b) defines impact as the effect that tests have on society at large, extending test influence beyond individuals or educational system, calling researchers' attention on political and ethical issues on score interpretation and test use. In recent years impact studies have attracted researchers' interests with the growing concern about the social influence of an assessment. Moreover, it has been acknowledged by many researchers that washback is within the scope of impact (e.g., Bachman & Palmer, 1996, 2010; Hamp-Lyons, 1998; McNamara, 1996, 2000; Shohamy, 2001). Washback takes place at the micro level of participants, mainly learners and teachers, while impact is beyond the instructional setting and used in wider educational context and social dimension.

Chapelle (1999) defines consequences as the “value implications of the interpretations made from test scores and the social consequences of test uses” (p.262), which presents a different dimension for a validity argument. Bachman and Palmer (2010) use the term consequence referring to test influence on various groups of stakeholders in their AUA framework. Their discussions about consequences for individuals including test takers and teachers, and consequences for educational systems and society indicate that they regard the concept and scope of consequence as similar to those of impact. Thus, Bachman and Palmer (2010) propose “washback should be considered within the scope of consequences. The term consequence tended to be used more frequently when linking test influence to validity and discussing consequential validity” (p.109).

The aforementioned part introduces definitions of washback and other related terms concerning test influence on teaching and learning, as well as test impact in social dimension. These terms will be remained as they are when quoting directly from the authors. However, when addressing influences of the CET-4, the study uses consequence interchangeably with washback in that the research background of the study is set within the instructional setting, and consequence is the terminology used in the AUA framewok. It should be noted defining washback only as effect of tests on teaching and learning is somewhat simplistic. What we present above is its general definition. The following part will discuss washback as a complex mechanism.

3.3.2 Theoretical washback studies

Along with the discussion of the concept of washback, specialists in language education also explore understanding of its nature, mechanisms and other aspects that are influenced by language testing. This part first discusses the direction and intentionality of washback. Then several theoretical models on washback mechanism are presented. Following this, major empirical washback studies are summarized to provide insights for methodology concerns of this study.

3.3.2.1 Nature of washback

Washback, regarded as neutral, can lead to both positive and negative effects (e.g., Alderson & Wall, 1993; Alderson & Hamp-Lyons, 1996; Bachman & Palmer, 1996; Bailey, 1996; Cheng, 2004; Qi, 2004). Wall and Alderson (1993) point out that tests can be powerful determiners, both positively and negatively, of what happens in classrooms. Bailey (1996) indicates that washback can be either positive or negative to the extent that it either promotes or impedes the accomplishment of educational goals held by learners and /or programme personnel.

Positive washback

Some language educators see washback in a positive way. Crooks (1988) believes that testing can have a positive effect on learning if teachers stress the need for ‘deep learning’, use evaluation to assist students, use feedback to focus students’ attention on their progress, set high but attainable standards, and select evaluation tasks to suit the goals being assessed. According to Herman and Golan (1993): public examinations set meaningful standards, provide feedback to improve classroom instruction, promote accountability of school systems, schools, and teachers for students’ learning, which can be used to enhance fast and broad changes within schools and thus to stimulate major educational reform. Alderson and Wall (1993) refer positive washback to tests and examinations that influence teaching and learning beneficially.

Negative washback

On the other hand, some language educators see language testing in a negative way. Negative washback is defined as the negative or undesirable effect on teaching and learning of a particular test (Alderson & Wall, 1993). Vernon (1956) claimed that examinations distort the curriculum since teachers tend to ignore subjects and activities that will not contribute to passing the exam. Pearson (1988) views the influence of public examinations on the attitudes, behavior, and motivation of teachers, learners, and parents as negative. Hughes (1989, p.1) worried that "if a test is regarded as important, then preparation for it can come to dominate all teaching and learning activities". Raimes (1990) discusses the potential negative washback of narrowing of the curriculum and proliferation of coaching and test-specific instructional materials. Prodromou (1995) argues that washback effect is "predominantly negative" (p.14) and is "one of the main reasons why new methods often fail to take root in language classes" (p.14). Messick (1996) indicates that negative consequence is tended to be linked to a source of invalidity. Brown (2000) holds that washback becomes negative when a mismatch emerges between the construct definition and the test, or between the content and the test. Quite a few of empirical studies have found that tests may affect teachers directly and negatively, imposing anxiety and accountability pressure on them (Fish, 1988; Noble & Smith, 1994), instructional time may be reduced by teaching testing-taking skills and drilling on multiple-choice items may boost scores but unlikely to promote general understanding (Noble & Smith, 1994).

The phenomenon of using tests to control the curriculum and shape teaching and learning is common in general and language education (e.g., Wall and Alderson, 1993; Spolsky, 1994; Cheng, 1997, 1998,). However, negative washback is more salient in high-stakes tests, which usually leads to measurement-driven instruction (MDI) referring to the notion that tests drive teaching and hence learning (Popham, 1987). Madaus (1988) warns that:

> The power of tests is a perceptual phenomenon. The higher the stakes attached to a test, the more it will distort the teaching process. Past exam papers eventually become the teaching curriculum. Teachers adjust their teaching to fit the form of exam questions. Test results become the major goal of schooling, and the agencies which set or control examinations eventually assume control over the curriculum. (p.88)

Alderson and Hamp-Lyons (1996. p.281) also list four main negative effects on curriculum that are often associated with 'high-stakes' language testing: 1)

Narrowing of the curriculum; 2) Test score pollution or increase in test scores without an accompanying rise in ability in the construct being tested; 3) Reduced emphasis on skills that require complex thinking or problem solving; 4) Lost instructional time.

To conclude, no consensus has been reached as to whether certain washback is positive or negative (Cheng & Curtis, 2004). Given the potentially bidirectional nature of washback, it should be recognized the consequences of language assessment, beneficial or detrimental, can be influenced by many factors. Bailey (1996, p.261) reminds "Positive washback is a primary goal for test developers", so it is the responsibility of test developers and test users that consider the extent to which any detrimental consequences may offset the intended beneficial consequences.

Investigating the nature of washback is only the first step. Bailey further suggests that there should be concerns about how to promote the former and inhibit the latter. Hughes (1989) outlines seven approaches to promoting positive backwash: 1) Test the abilities whose development you want to encourage; 2) Sample widely and unpredictably; 3) Use direct testing; 4) Make testing criterion-referenced; 5) Base achievement tests on objectives; 6) Ensure test is known and understood by students and teachers; 7) Where necessary provide assistance to teachers. Bailey (1996) later offers some criteria from her thorough review of the existing literature to promote beneficial washback: aligning instructional content with educational goals, increasing authenticity of test tasks, introducing learner autonomy and self-assessment, providing detailed score reporting. Messick (1996) affirmed that "for optimal positive washback there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test" (p.242). As the present study is set within the reform background to investigate the CET-4, reviews on washback nature and on approaches to promote positive washback provide useful anchors to better evaluate and understand rationales behind the CET-4 reform and the theoretical underpinnings for the reformed test tasks.

3.3.2.2 Mechanism of washback

Messick (1995) suggested "What matters is not only whether the social consequences of test interpretations and use are positive or negative, but how the consequences came and what determined them" (p.748). Washback is also viewed as complex and multifaceted, interweaved with a great many variables (e.g., Cheng, 2005; Hawkey, 2006). Thus, in this part, I will briefly review models proposed to exploring how washback works or functions.

Alderson and Wall's washback hypothesis

Following Messick's concern on consequences of test use, Alderson and Wall(1993) set out their theoretical and empirical studies on washback, and propose fifteen washback hypotheses:

> 1) A test will influence teaching.
> 2) A test will influence learning.
> 3) A test will influence what teachers teach.
> 4) A test will influence how teachers teach.
> 5) A test will influence what learners learn.
> 6) A test will influence how learners learn.
> 7) A test will influence the rate and sequence of teaching.
> 8) A test will influence the rate and sequence of learning.
> 9) A test will influence the degree and the depth of teaching.
> 10) A test will influence the degree and the depth of learning.
> 11) A test will influence attitudes to the content, method, etc. of teaching and learning.
> 12) Tests that have important consequences will have washback.
> 13) Tests that do not have important consequences will have no washback.
> 14) Tests will have washback on all learners and teachers.
> 15) Tests will have washback effects for some learners and some teachers, but not for others. (p.120-p.121)

The above hypotheses fall into three categories. Hypotheses 1), 3), 4), 7), 9), 11) are concerned with washback on teaching, containing facets of teaching content, teaching methods, and teachers' attitudes, while hypotheses 2), 5), 6), 8), 10), 11) are related to washback on learning involving facets of learning content, learning strategies, and learners' attitudes. The remaining hypotheses 12), 13), 14), 15) deal with the strength or intensity of washback, varying with test stakes and test takers' characteristics. Alderson and Wall's work provides food for thought to a series of subsequent theoretical and empirical washback studies, and to date has been viewed as most comprehensive washback literature in language testing field.

Hughes' Basic Model of Backwash

Hughes (1993) clearly distinguishes between participants (students, classroom teachers, administrators, material developers, publishers), process (material developments, syllabus designing, changes in teaching methodology, the use of learning and / or test-taking strategies, etc), and product (what is learned such as facts and skills) in his basic model of backwash. Hughes's (1993) further discusses how washback works by his trichotomy mechanism:

> The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. Their perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work. (p.2)

Hughes' model shares similarities with Alderson and Wall's hypotheses in that they all distinguish between participants and processes. What differs is that "the former deals with the scope and pattern of the whole educational system that comes under the influence of tests whereas the latter focuses on the effect on teaching and learning" (Qi, 2004, p. 39).

Bailey's washback model

Based on the combination of Alderson and Wall's hypotheses and Hughes' trichotomy model, Bailey (1996) classifies washback into two categories, washback to the learners, and washback to the program, depending on the participants we consider. The former mainly refers to effects on test takers while the latter involves effects on teachers, administrators, curriculum developers, counselors, etc. Likewise Bailey's Model share similar aspects with Alderson and Wall's washback hypotheses, among which hypotheses 2), 5), 6), 8), 10) are in line with washback to the learners while hypotheses 1), 3), 4), 7), 9), 11) are corresponding to washback to the program (Bailey, 1996, p.265). Bailey (1996) further summarizes processes students may participate in when faced with an important test:

> 1) Practicing items similar in format to those on the test.
> 2) Studying vocabulary and grammar rules.
> 3) Participating in interactive language practice (e.g., target language conversations).
> 4) Reading widely in the target language.
> 5) Listening to noninteractive language (radio, television, etc.).
> 6) Applying test-taking strategies.
> 7) Enrolling in test-preparation course.
> 8) Requesting guidance in their studying and feedback on their performance.
> 9) Enrolling in, requesting or demanding additional (unscheduled) test-preparation classes or tutorials (in addition to or in lieu of other language classes).
> 10) Skipping language classes to study for the test. (p.264-p.265)

Bailey indicates that selection among these processes would lead to either beneficial or negative washback on the basis of the criterion whether their use promote the learners' actual language development.

To conclude, Alderson and Wall's Hypothesis reveal the interrelationship between testing, teaching, learning, and the interactive effect on teachers and learners. Their hypotheses theoretically undergird the present study. Given the wide acknowledgement and consensus that washback does exist, hypotheses related to washback on teaching and learning will be investigated in the present study, which are in accordance with the third research question proposed in Chapter 1 (see section 1.3, p.8). Alderson and Wall (1993) also suggest that further research on washback should incorporate findings in the areas of motivation and performance, as well as educational innovation. This inspires me to embrace these facets into my research design. Hughes's washback model expands the influence of tests on teachers and learners to broader range such as researchers, material writers and curriculum designers. His model provides rationale for investigation into both teachers' and students' perceptions of and attitudes to the CET-4. Their perceptions may affect their teaching and learning activities and behaviors, which in turn may have an effect on students' test performances, so students' test performances will be linked to washback phenomenon. The above models provide insights to break down the third RQ into more specific and operational research questions, which will be discussed with the consequence claim and the washback warrants to be articulated in Chapter 4 (see Section4.4.4, p.110). The processes proposed by Bailey also provide insights for the questionnaire designing of the present study. Items can be adapted from the processes to examine student's learning behaviors in their test preparation activities.

3.3.3 Empirical washback studies

As discussed earlier, Alderson and Wall's (1993a) study explores the concept of washback and addresses a series of hypotheses. In spite of lack of empirical evidence, their study has made great contribution to inspiring language testers' interests in washback issues and providing insights for subsequence research. However, it has to be admitted that a large number of theoretical studies are based on assertions and reported perceptions rather than direct classroom observations. Therefore, empirical studies have been conducted since the late 1990s to explore washback phenomenon more extensively and profoundly. This section will review some influential empirical studies.

Alderson and Wall (1993b) examined washback of a new O-level English Test introduced in Sri Lanka. The prominence of this study is the adoption of classroom observation approach in addition to interviews. They found that the impact of the new examination was less pervasive than that had been expected. Although the examination had negative washback on teaching content, narrowing the curriculum, it had basically no impact on teaching methodology.

Alderson and Hamp-Lyons (1996) examined effects of TOEFL preparation classes in the US. They conducted questionnaires, group interviews and classroom observations with comparison groups, and found that TOEFL affects both teaching content and method, but washback intensity varies with individual characteristics of teachers.

Shohamy, Donitsa-Schmidt and Ferman (1996) investigated the long-term effects of two national tests in secondary schools in Israel, one in Arabic as a second language test and one in English as a foreign language test through questionnaires, interviews and document analysis from teachers, students and inspectors. They found that washback varies over time due to many factors such as the language status and uses of the test.

Watanabe (1996) studied washback effects of the introduction of a translation test into the university Entrance Examination in Japan. Watanabe conducted classroom observations and interviews with two teachers. This small-scale study revealed that teacher factors, such as educational background, personal beliefs, and teaching experience may outweigh the possible effect of the entrance examinations.

Cheng (1997, 1999, 2005) examined the impact of a revised high-stakes examination, the 1996 Hong Kong Certificate of Education in English, on the classroom teaching of English in Hong Kong secondary school. The studies focused on teachers, exploring the classroom levels of teaching and learning. Cheng employed both quantitative and qualitative methods including questionnaires, interviews, and classroom observations to investigate effects on teachers' attitudes, teaching content, and classroom instruction. One of merits in her studies is that she conducted a baseline study to compare the test effects before and after the introduction of the new test. Her studies found that the new test brings about changes in teaching materials and contents but limited changes in methods teachers employ.

Burrows (2004) conducted a study on washback in classroom-based assessment in the Adult Migrant English Program in Australia. Her study combined classroom observations with teacher interviews, and focused on teachers' beliefs, and attitudes to a new test. She found that teachers' attitudes to a new test will have an effect on the test implementation.

Qi (2004, 2005, 2007) focused her studies on the writing tasks of the National Matriculation English Test (NMET) in China, exploring whether a test can affect teaching in the way intended by test developers. She collected data from an extensive sample of stakeholders of test constructors, teachers and students via interviews, questionnaires and classroom observations. She found that high-stakes tests are not an efficient agent for pedagogical change. The selection function and the function of promoting change impeded it to achieve the intended washback.

Saville and Hawkey (2004) investigated the impact of IELTS on teaching materials by conducting teacher surveys and analyzing textbooks and test practice books. They found that teachers tend to base their teaching materials evaluation on the extent to which the materials reflect test content and format.

Green (2007) investigated the influence of IETLS Academic Writing Module on preparation for academic study and equivalence between IELTS test preparation and other forms of English for Academic Purposes directed at university study. His study focused on students and incorporated test taker factors like their perceptions of the test, expectancy of test demand, and the value they placed on test into his washback model. Data were collected through a comprehensive range of methods including questionnaires, interviews, classroom observations, test instruments and document analysis. His work provides a valuable framework exploring the relationship between test scores, periods of studies and language gain.

Wall and Horák (2008) explored teachers' awareness of and reactions to the changes in the new TOEFL. They adopted computer-mediated interviews to collect data. They found that teachers had low awareness of the changes at the initial stage of their study, but with their awareness developing, they were found positive about the new TOEFL. Wall and Horák suggest that it should be of importance to disseminate information about the test in order to bring about the intended impact of a new test.

In summary, the aforementioned part reviewed several influential and illuminating empirical studies on washback. It can be seen that washback is far more complex and unpredictable, affected by intervening variables like teacher and student related factors, status of language, test uses, teaching and test preparation materials, etc. Second, most studies are conducted to explore test influence on teachers while test influence on learning and learners is scarcely researched. Third, questionnaires and interviews are the major instruments adopted to explore stakeholders' attitudes, while classroom observation has been more frequently adopted to explore what actually happens in classrooms. In addition, test instruments are employed linking test takers' performances to their perceptions, motivations, preparation activities, etc. to explore factors contributing to score gains and factors influencing how washback really works. Finally, although washback studies mentioned above are grounded in theoretical rationales and certain operational framework or mechanisms, they are basically discussed separately from validity.

The above discussions also help identify gaps and offer some insights for the present study. The study will focus more on students but include stakeholders of teachers and administrators from the University Academic Affairs Offices as well. Students' and teachers' attitudes to certain aspects of washback will be compared. In addition, student's test performance will be linked to their perceptions. Multifaceted aspects of washback are to be explored via some specific questionnaire items. The detailed statistical analyses and research design to achieve the above ends will be elaborated in Chapter 6. Above all, this study will draw on an AUA framework to link washback with validity issues. The following part will touch upon the role of consequences in validation.

3.3.4 The role of consequence in validation

This part addresses the role of consequence in validation. It first traces the role of consequence in the early validity models, and then examines its function in recent validity frameworks.

As indicated earlier, during the early 1950s to the late 1970s, validity had been mainly viewed as a toolkit in which the Trinitarian model of criterion, content and construct validity was widely embraced to collect evidence to support the soundness of interpretations and use of test scores (Cronbach, 1988; Guion, 1980; Messick, 1989). The approach of using different models as a toolkit in validation is problematic for lacking a coherent and logical argument to guide the process of validation (Bachman, 2005; Kane, 2006, 2012). This historical view of types of validity left almost no place for consequences of test score use. Interpretations of test scores and uses of test scores were treated separately. The consequences of test score use were not considered as part of evidence for arguments supporting test score interpretation (Nichols & Williams, 2009).

The growing concern on consequences of test use can be traced back to Messick's (1989) introduction of his unified validity framework into which he incorporates notions of value implication and social consequence. Messick (1995) suggested that the consequential aspect of construct validity, or consequential validity, should include evidence and rationales for evaluating the intended and unintended consequence of score interpretation and immediate and long-term test use. To his credit, Messick's validity matrix made consequence from an implicit aspect of validity to an explicit and prominent component of validity evidence (Kane, 2006; Nichols & Williams, 2009), sparking language testers' strong interest in investigating consequential aspect of validity. However, in spite of the solid rationales Messick's work has provided for the subsequence theoretical and empirical impact and washback studies, Messick failed to provide operational procedures to investigate them.

Since Messick's introduction of his unified validity framework, there has been much debate on whether test consequences should be part of validity concept. Some researchers argued that concept of validity should mainly involve the descriptive interpretation of scores, and validation should be separated from consequence of test use (Mehrens, 1997; Popham, 1997). They argued that incorporation consequence of test use into the realm of validity not only made the validity concept more complex, but also burdened test developers with the accountability to collect evidence of consequence of test use, especially when investigation into the inference of score interpretation of a test construct was independent of any specific use of the score (Green, 1998; Mehrens, 1997; Pomplun, 1997, Reckase, 1998).

Other researchers contended that consequence of test use and score interpretation are an integral aspect of validity (e.g., Cronbach, 1988; Linn, 2005; Messick, 1975, 1980, 1989, 1996; Moss, 1992; Shepard, 1997). Cronbach (1988) suggested that it was essential to embrace consideration of consequence when evaluating the legitimacy of test use. Shepard (1997) viewed test consequence as a logical part of test validation and called for more emphasis on the intended test uses. Linn (2005) agreed with Shepard that conception of validity should be expanded to include consequences of test use.

In terms of washback and validity, Morrow (1986) coined the term washback validity, claiming that "The first validity criterion that I would …put forward would be a measure of how far the intended washback effect was actually being met in practice" (p.6). Frederiksen and Collins (1989) introduced the term systemic validity, similar to washback validity. Alderson (1995) further noted "washabck is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test" (p.3). Messick (1996) echoed "evidence of teaching and learning effects should be interpreted as washback…only if that evidence can be linked to the introduction and use of the test" (p.252). In addition, Messick (1996) stressed that washback is only one form of testing consequence that needs to be weighed in evaluating validity, and testing consequences are only one aspect of construct validity needing to be addressed.

The inclusion of consequences in the current discussion of validity evidence has been a contentious issue for a long time. However, broad consensus has been reached today that the consequences of test use have implications to score interpretations. These implications provide meaningful and important sources of validity evidence when consequences can be linked to construct underrepresentation or construct irrelevant variance (e.g., AERA, APA, & NCME, 1999; Cronbach, 1988; Kane, 2006; Messick, 1989, 1996; Nichols & Williams, 2009).

More recently, the consequence of test use has begun to receive increasing attention in the argument-based approaches to validation. Validity is defined as judgments of the degree to which arguments support the interpretations and uses of the test scores An investigation into test consequences is viewed as a justification of test uses, which is part of the test validation process (Bachman, 2005; Bachman & Palmer, 2010; Kane, 2006, 2012; Nichols & Williams, 2009). Kane (2002, 2006) embraces test consequences as an important component in his interpretative argument and further delineates his interpretative argument as descriptive and decision-based interpretations. Bachman and Palmer (2010) incorporate a decision claim as a coherent and logic inference in their AUA framework to link interpretations of test scores to consequences of test uses. It has been acknowledged that various uses of the scores should be based on the proposed interpretations in evaluating the appropriateness of decisions. In high-stakes tests, cut scores are usually employed to make decisions. In order to avoid classification decision errors and minimize the induced negative consequences, performance standard should be specified first with verbal descriptions of achievement level. Then, a cut score can be set based on corresponding performance standard and well-developed technical procedures (Bachman &Palmer, 2010; Kane, 2006, 2012).

The discussion in this regard continues by focusing more attention on the role of social consequences in validation. Kane (2012) stresses investigating kinds of social consequences, approaches to evaluating these consequences, and people held accountable for such evaluations. As discussed earlier, adverse social consequences were traditionally traced to sources of construct underrepresentation or construct-irrelevant variance. However, evaluations of social consequences under validity have expanded over time. Societal values, educational system, policy issues, legal regulations, and equitability serve as important concerns in the score-based decisions that are made (Bachman, 2005; Bachman & Palmer, 2010;

Kane, 2012; McNamara & Roever, 2006). In addition, the role of test developers and test users in collecting evidence to support use of test scores is more clearly defined. Test developers should be responsible for the claims they made. They have the responsibility for delineating the intended test purpose, test uses and the expected beneficial consequences to be brought about, informing test users and test takers of any unintended test effects as well as minimizing the side effects (e.g., Bachman & Palmer, 2010; Kane, 2006, 2012; Moss, 1998; Shepard, 1997). With regard to test users, Kane (2006, 2012) maintains that they are responsible for collecting evidence to support uses of scores they proposed in the social dimension. Test developers are not supposed to take responsibility for the negative consequence resulting from any misuse of test scores.

Section summary

This section has reviewed literature on washback and its mechanism. This strand of theoretical underpinnings provides insight into the design of questionnaire surveys. Above all, the discussion on relationship between washback and validity helps identify the research gap and reveal limitations of previous studies. Viewing washback as part of validity is only the initial step. The next step is to establish logical and coherent inferential links from interpretations of test scores to test influence. Thus, washback review in this section also provides underpinnings for drawing this study on the AUA framework.

3.4 Summary

In this chapter, a historical approach has been adopted to present the transformations of validity both in its scope and in concept since its introduction about sixty years ago. The evolution of test validity concept has been delineated from classification of types of validity to construct validity as an overarching concept. It is now widely accepted that validity is associated with the interpretations assigned to test scores rather than with the scores themselves or the test and involves an evaluation of the appropriateness of these interpretations. In the late 1980s, Messick’s (1989) unified validity framework advanced validity theory by incorporating value implications and social consequences into validation framework. Flaws of Messick's work lie in the abstractness and failure to provide guidance and procedures to conduct the real validation research. Since the 1990s much research has emerged addressing the consequences of test uses, esp. washback of high-stakes tests, but they separated washback from validity(e.g., Alderson & Hamp-Lyons, 1996; Alderson & Wall,1993; Andrews, 1994; Cheng,1997, 1998; Gu, 2004; Jin, 2000; Messick, 1996; Qi, 2004; Wall, 1996; Watanable, 1996). Other researchers have attempted to link validity to test uses, consequences and ethical considerations, but failed to establish an explicit and coherent linkage among those qualities or between validity and test use (e.g., Bachman & Palmer, 1996; Kunnan, 1997, 2004; Schohamy, 1993, 2001). More recent argument-based formulations of validity provides explicit and logical links between test takers' performance and score-based interpretations, but are unclear about how these relate to assessment use — decisions and consequences e.g., Kane, 1992, 2001, 2002, 2004, 2006; Kane et al., 1999; Mislevy, 1996, Mislevy et al., 2002, 2003). Therefore, given limitations of previous studies, current concern about issues of consequences, ethics, and fairness in language assessment, and the increasing recognition of the need to link validity issues with the consequences of using language tests, Bachman and Palmer, fusing Toulmin's argument model with Messick's unitary view of validity, proposed an AUA, an all-encompassing framework to date. An AUA establishes logical and coherent inferential links from test performances to interpretation of scores, from score-based decisions eventually to consequences of an assessment. This study is grounded in Bachman and Palmer's AUA framework, attempting to build a locally specific AUA for the CET-4.

The second part of this chapter has touched upon washback, addressing its concept, nature, mechanism, and methodological approaches. Review of this part is expected to provide a basic background of washback as a complex and multifaceted concept. It also contributes to identifying research focus and research design of this study. Thus, both teachers' and students' perceptions of the CET-4, their teaching and learning behaviors and test preparation activities will be examined. Relationship between test performances and test takers’ perceptions and learning behaviors are to be explored as well.

These two strands of literature view provide justifications of drawing the present study on an AUA framework as well as theoretical underpinnings for claims and related warrants to be articulated in Chapter 4. In addition, insights are also generated for the methodological concerns to be discussed in Chapter 6.

CHAPTER 4

ARTICULATION OF AN AUA FOR THE CET-4

4.1 Introduction

In Chapter 1, three research questions are proposed in relation to the overarching research purpose. In Chapter 3 a thorough review of the structure, elements and qualities that constitute an AUA is conducted, which provides a theoretical foundation for articulating an AUA for the CET-4. In this chapter, the three research questions are first linked to the corresponding claims. It follows a justification of why these claims can be articulated to evaluate the CET-4. The chapter finally delineates how specific and relevant warrants or rebuttals of an AUA are articulated in the CET-4 context.

4.2 Linking research questions to claims

This study is intended to investigate in what way and to what extent the CET-4 can serve as a useful indicator of students' overall English proficiency and an effective measure to promote English teaching and learning. Corresponding to this overarching research purpose, three research questions related to validity, score-based decisions, and washback issues of the CET-4 were proposed in Chapter 1. The first step in this chapter is to discuss the research questions in relation to Bachman and Palmer's AUA framework and to establish links between them so that a local AUA consisting of specific claims, warrants or rebuttals can be articulated within the context of the CET-4 as a conceptual framework to guide our research objectives.

Figure 4.1 illustrates the links between research questions and the claims. The left column presents the generic claims in the AUA template produced by Bachman and Palmer (2010, p 158), while the right column presents the three research questions. The single direction arrow pinpoints their corresponding relationships.

The inferential link in an AUA framework can be bidirectional. With regard to assessment development, a top-down order is advised to start with a consequence claim. In contrast, interpreting test scores and justifying test uses prefers a bottom-up order. Since the research questions start with interpretation of scores to decisions made on test scores, and to test consequences, a bottom-up order is followed in the present study.

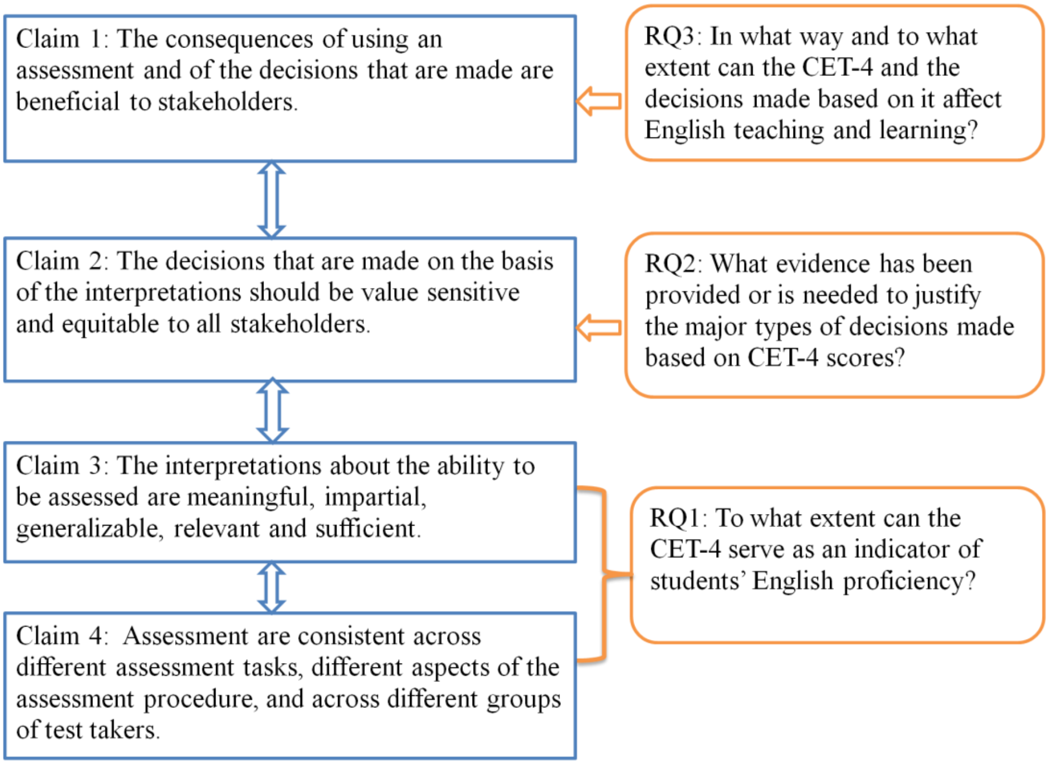
Claim 1: The consequences of using an assessment and of the decisions that are made are beneficial to stakeholders.
RQ3: In what way and to what extent can the CET-4 and the decisions made based on it affect English teaching and learning?
Claim 2: The decisions that are made on the basis of the interpretations should be value sensitive and equitable to all stakeholders.
RQ2: What evidence has been provided or is needed to justify the major types of decisions made based on CET-4 scores?
Claim 3: The interpretations about the ability to be assessed are meaningful, impartial, generalizable, relevant and sufficient.
RQ1: To what extent can the CET-4 serve as an indicator of students' English proficiency?
Claim 4: Assessment are consistent across different assessment tasks, different aspects of the assessment procedure, and across different groups of test takers.

Figure 4. 1 Links between research questions and claims in an AUA (adapted from Bachman & Palmer, 2010, p.158)

RQ1 is related to the validity and reliability of interpretations of test takers' CET-4 scores, so it should be linked to the claims of assessment records and interpretations. RQ2 aims to examine the major types of decisions that are made based on CET-4 scores awarded to test takers and the evidence justifying these decisions, which is directly linked to claim of decisions. The decision claim serves as an indispensable inferential link bridging interpretations of score meanings with test consequences. RQ3 seeks to investigate in what way and to what extent the

CET-4 and the decisions made on it affect English teaching and learning. This question mainly discusses washabck of the CET-4, which is linked to the consequence claim on test uses. Figure 4.1 presented above is only a schematic graph indicating the links between research questions and the generic claims provided by Bachman and Palmer (2010). Before the articulation of a specific AUA for the CET-4, it is necessary to justify why these claims can be applicable to evaluate the CET-4 and explain how Bachman and Palmer's AUA framework is adapted to the CET-4 in Chinese assessment setting.

4.3 Justification of the claims on the CET-4

Before the articulation of an AUA for the CET-4, issues about legitimacy of claims on the CET-4 need to be addressed first. As noted earlier, up to now, few studies have been conducted by adopting an AUA as a framework either to develop an assessment or to justify its uses. Admittedly, the CET-4 is not developed based on an AUA, nor does the NCETC make the following claims in terminologies used by an AUA. Thus, it is essential to first clarify questions or doubts about why the study draws on the claims in an AUA to evaluate the CET-4 and to guide backing collection whilst the NCETC did not make such claims. The following part explains why the articulated claims on the CET-4 in this study can hold up both at theoretical and operational dimensions.

Given its large test population and far-reaching influences, the CET-4 has been acknowledged as a large-scale and high-stakes test (e.g., Cheng, 2008; Jin, 2005, 2006, 2008, 2010; Liu & Dai, 2004; Wang, 2010; Wu, 2005; Yang & Weir, 1998; Zhao & Cheng, 2010; Zheng & Cheng, 2008). It follows that this test is supposed to incorporate qualities that large-scale and high-stakes tests are expected to possess. Bachman and Palmer (2010) explicitly state that the major purpose of using an assessment is to collect information that is used to make decisions, and the uses of an assessment and the decisions that are made have consequences for different groups of stakeholders. I would like to refer to the three tests (TOEFL, ITELS, NMET) mentioned in Chapter 1 to facilitate my justification, since they are all categorized as large-scale and high-stakes tests.

ETS (Educational Testing Service) initiated TOEFL 2000 projects and sponsored a large body of research (e.g., Alderson & Hamp-Lyons, 1996; Chapelle, Enright, Jamieson, 2008; Cohen & Upton, 2006; Sawaki, Stricker & Oranje, 2008; Wall & Horak, 2006, 2007, 2008) to investigate reliability, validity and washback of a new TOEFL test, which is the subsequent TOEFL iBT (Test of English as a Foreign Language: internet-based Test) officially launched in 2005. Based on solid evidence from these studies, ETS makes explicit statements about reliability and comparability of TOEFL iBT scores across test forms (ETS, 2010), as well as its score meaning, decisions that are made, and test consequences it is intended to bring about. TOEFL iBT scores are interpreted as test takers' ability to use and understand English in college and university settings, and are used to make admissions and placement decisions. The aim of the TOEFL iBT test is to maximize the positive consequences of score use (ETS, 2011, p.3, p.9).

Likewise, similar claims about IELTS can be observed on the IELTS website (www.ielts.org) which lists both completed and ongoing research projects and reports its recent test data. The extensive research on IELTS makes an important contribution to monitoring the rigorous process to produce the test materials and ensuring the test being fair and unbiased in its continued development and improvement (e.g., Clapham, 1996; Moore & Morton, 2007; Moore, Morton & Price, 2012; Taylor, 2001). Thus, the Cambridge ESOL (the University of Cambridge English for Speakers of Other Languages) maintains that IELTS is recognized as a secure, valid and reliable indicator of true-to-life ability to communicate in English for education, immigration and professional accreditation (IELTS, 2010). The rigorous processes used to produce the test materials ensure that every version of the test is of comparable level of difficulty so that candidates' results are consistent wherever and whenever they take the test (IELTS, 2010) In addition, impact and washback have become an increasing concern for the Cambridge ESOL. Quite a number of studies have been conducted with the purpose to promote beneficial impact and washback of IELTS on language learning and teaching (Green, 2007; Hawkey, 2006; Saville & Hawkey, 2004).

The NMET, which apparently shares the same testing context with the CET-4, can further underlie my argument here. The purpose of NMET is to make inferences about candidates' English language ability, which is used for university admissions decisions. In addition to its selective function, the NMET is also intended to bring positive washback to English teaching and learning at the secondary level (Cheng, 2008; Cheng & Qi, 2006; Qi, 2010). To achieve its dual purposes, the NMET constructors take a range of measures to ensure test qualities such as validity, reliability, discrimination power, fairness, and positive washback of the test on teaching and learning (Li, 1990; Qi, 2004, 2005, 2010).

Based on the discussion of the three tests, it can be seen that a large-scale test is expected to have a clear definition of its construct, and explicitly informs its test takers and users of what the test is intended to measure, and what score based decisions are to be made. In addition, given its far-reaching influence and high-stakes nature, information about test takers' language proficiency should be generated from reliable test scores and valid score interpretations. Beneficial consequences should always be a great concern to test agencies and test developers. It follows that the CET-4, with its over 10 million test takers per year and is supposed to assume similar qualities as well. Evidence can also be observed from statements on the CET official website (www.cet.edu.cn): The purpose of this test is to objectively and accurately measure students' English proficiency. The CET possesses higher reliability and validity. The test difficulty is comparable and its scores are consistent. It promotes the implementation of the NCETS and improvement of College English teaching and learning. Thus, both the inferences from discussions on the three tests above and the statements made by the NCETC on the official CET website provide the legitimacy to apply these claims to evaluate the CET-4.

A review of empirical studies on evaluation of the CET-4 can also provide support for the adoption of an AUA. In the CET validation study conducted by a three-year Sino-British joint research project (Yang & Weir, 1998), the NCETC followed the traditional concept of validity to investigate the construct validity, content validity, concurrent validity, predictive validity and face validity of the CET. Later Zheng and Cheng (2006) draw on Kunnan's test fairness framework to appraise the CET-4 according to the five qualities of validity, absence of bias, access, administration, and social consequences. These studies suggest that it is feasible to employ different frameworks to evaluate the CET-4 even though the test per se is not designed based on a given framework. Thus, drawing on the AUA to examine the CET-4 is a tenable approach and is expected to provide a fresh perspective on evaluation of the CET-4.

4.4 Adapting the generic AUA template to the CET-4 scenario

Bachman and Palmer call for attention to several points in articulation of an AUA. They reiterate that what they provide is only a generic structure and terminology of an AUA, which are illustrative rather than prescriptive. Hence, claims, warrants and rebuttals should be stated in terms that apply specifically to an assessment. In addition, not all the warrants and rebuttals illustrated in their template and examples will be required in the AUA for a given assessment. Finally, although for every warrant there is an implied rebuttal, it makes more sense to argue for warrants rather than to state them as implied rebuttals (Bachman & Palmer, 2010). An AUA is a conceptual framework to guide the process of test development or justification, but it does not mean this framework has to be strictly followed and replicated without adaption. Rather, it should be used flexibly to make the AUA locally specific. Therefore, considering Chinese testing setting and practical constrains of this study like data access, time and money resources, adaption is made in the process of articulating the AUA for the CET-4.

In respect to Bachman and Palmer' AUA framework presented by Figure 3.4 in Chapter 3, I have explained that the downward pointing arrow on the left signifies the process of assessment development while the upward pointing arrow on the right signifies the process of interpretation and use. In other words, when developing an assessment, a top-down approach is preferred, because beneficial consequences of using the test should always be the prioritized concern. On the contrary, justification of assessment uses should start from the bottom. Since the present study aims at investigating uses of the CET-4, a bottom-up approach is adopted, which means the claims in Bachman and Palmer's AUA framework will be presented in a reverse order. Thus, a chain of inferences will start from test takers' performances to assessment records, from interpretations to decisions, and to consequences of using the CET-4.

4.4.1 Assessment records

Claim: The CET-4 scores are consistent across different assessment tasks and administrations of the test, and across different groups of test takers.

Based on the generic structure of claims in Bachman and Palmer's AUA framework, the consistency claim about the CET-4 is articulated in the above box. Bachman and Palmer (2010) produce ten illustrative warrants under the consistency claim. These warrants involve aspects of test administration procedures, scoring criteria, rater training and reliability, and comparable consistency of scores. I will first explain why warrants related to the CET-4 to support this claim are not articulated here. The present study is not intended to organize a mock CET-4 administration. Test data will be collected from the official CET-4 Score Report Forms with test takers' overall score and the scores of each component. In spite of my efforts to contact the NCETC, I did not obtain the detailed score profiles of samples in my study, such as their item scores and subsection scores of listening and reading components. It means advanced statistical measures such as classical item theory and Rasch model cannot be performed to investigate the internal consistency of reliability, comparability of scores, and rater consistency. Since I have no access to the central database of CET scores, which are needed for supporting consistency claims and warrants, I have to accept what evidence is available to me. The evidence obtained includes documentation provided by the NCETC on procedures of test administration, scoring and reporting, statements made by the NCETC and findings released in their studies, which are briefly summarized in the following part.

As mentioned in Chapter 2, the National Education Examinations Authority and its subordinates at provincial or municipal level are in charge of administration of the CET, including registration, test delivery and coordination of the marking centers (Jin, 2010). Detailed and clear administrative procedures have been laid down to ensure the rigor and fairness of the test (Jin & Yang, 2006).

In terms of rater training and rating reliability, the NCETC declares to have established a qualified team of essay markers and a strict system of quality control (Jin, 2010). Chapter 2 has introduced the CET essay marking system and quality control procedures. Range-finders and marking schemes are provided for rater training and the final writing score is subject to computer adjustment to filter out inconsistencies resulting from marker subjectivity (Jin & Yang, 2006; Yang 2003). Supervisors and group leaders are jointly responsible for spot-checking of the quality in each marking center. Research showed that the CET marking reliability is 0.87 (Jin, 2010).

The NCETC has taken a series of measures to meet the professional quality requirements of a large-scale standardized test. With respect to statistical analysis and score reporting, a series of computer procedures have been developed for machine reading, IRT equating, writing score adjustment, and score normalization, thus ensuring objectivity and consistency in marking and scoring, comparability and interpretability of test results, fairness of test administration conditions, rigor and efficiency in test administration (Jin & Yang, 2006).

To sum up, this section first explains why the consistency claim about the CET-4 is not to be investigated in this study. Then it reviews findings and conclusions made by the NCETC as backing evidence to consistency of scores. Despite our acceptance of the NCETC's statement on consistency of scores, it should be noted that both TOEFL iBT and IETLS post research reports on their test qualities, analysis of scores, guides for different test users on their websites as evidence of reliability and validity of their tests. In comparison, the information and evidence revealed by the NCETC publicly is limited and lack of transparency. It may be a potential source of rebuttal to the qualities of the CET-4. Given the above reasons, the study concentrates on the other three claims. I begin with claim of interpretations, which is marked as claim 1 in the present study, corresponding to the first research question.

4.4.2 Claim on interpretations

Claim 1: The interpretations of students' overall English proficiency are meaningful with respect to the uniform CECR and the CET-4 Syllabus and classroom teaching and learning activities, fair to all the test takers, generalizable to TLU and language teaching domains, relevant to and sufficient for the decisions that are to be made.

Based on the generic version offered by Bachman and Palmer (2010), the claim on interpretations of CET-4 scores is articulated in the above box in response to RQ1. The interpretation claim entails qualities of meaningfulness, impartiality, generalizability, relevance and sufficiency. Each quality can be supported by several warrants. Bachman and Palmer suggest that it is unnecessary to articulate all the illustrative warrants they list. Given practical constraints and limited access to test data, it is also impossible for the present study to engage all the warrants as a test agency can do. Thus, any warrants to be articulated should take into consideration the research questions and the specific CET-4 context. Since validity is an umbrella concept subsuming multifaceted aspects, RQ1 is broken into two specific and operationally defined sub questions:

RQ 1.1 To what extent does the CET-4 measure the construct to be assessed?

RQ1.2 To what extent is the CET-4 representative of the content relevance and coverage in accordance with the test syllabus and curriculum objectives?

In line with the two operationalized questions, three warrants are articulated below.

Warrant 1: The constructs to be assessed by the CET-4 are based on a frame of reference such as the uniform CECR, the CET-4 Syllabus, a needs analysis, or current theory of language use.

Warrant 2: The CET-4 can be interpreted as a useful indicator of the ability to be assessed.

Warrant 3: The CET-4 is meaningful and generalizable for its content representativeness and content relevance in accordance with the test syllabus and curriculum objectives.

Warrants of meaningfulness involve major aspects of construct definition, task specifications, administration procedures, scoring procedures, and the real test taking process. Impartiality is established on the fact that test contents should not favor any group of test takers and be free from any bias. Individuals should also have equal access to information about assessment content and procedures, and equal opportunity to prepare. In Chapter 2 when I introduce general and specific features of the current CET-4 (see section 2.4.4), I have discussed the test administration environment, test rubric, and task specifications delineated in the teaching and test syllabuses as well as scoring keys, marking procedures, and the Score Report Form. The CET-4 Syllabus clearly specifies what kinds of tasks will be presented to test takers and how they will be expected to respond to these so that inferences can be made from their performances about the constructs to be assessed. Thus, warrants related to test construct will be our focus. Justifications of the constructs can be conducted by statistical analysis of test scores like correlations and EFA. In addition, qualitative evidence can be collected via content analysis. While analyzing listening and reading components, judges will be asked to report any prejudiced and offensive contents. Students will also be asked about the issues relevant to impartiality in group interviews. However, it has to be admitted that given lack of item scores more advanced statistical procedures like differential item functioning (DIF) or confirmatory factor analysis (CFA) cannot be conducted to investigate whether the interpretations have equal meaning across different groups of test takers.

Generalizability means that the characteristics (e.g., the setting, rubric, input, expected response, and relationship between input and expected response) of the assessment tasks closely correspond to those of TLU tasks and the instructional tasks (Bachman & Palmer, 2010). Unlike TOEFL iBT and IELTS (academic module) which explicitly define their TLU domain as academic English use in campus, the CET-4 is designed as an achievement test for two years' compulsory college English teaching, whose objective is to enable students to use English in their future work and social interactions. It can be estimated that the TLU domain of the CET-4 is characterized by breadth and diversity, which may account for why little data or evidence about generalizability of the test can be traced from the NCETC projects or studies of individual researchers. To deal with this dilemma, I integrate generalizability with the meaningfulness warrant and investigate it with reference to teaching and testing syllabuses, since both address authenticity as the principle to select listening and reading materials. In addition, some items will be designed in questionnaires to explore stakeholders' opinions of test content such as authenticity, communicative features of test tasks, which is expected to provide additional backing for generalizability.

4.4.3 Claim on decisions

Claim 2: The multiple decisions that are made on the interpretation of the CET-4 scores reflect the existing educational and societal values and the relevant university regulations, and are equitable for all the stakeholders to be affected by the decisions.

RQ2 investigates multiple uses of the test and serves as an essential bridge to link interpretations of scores to consequences of test use. Based on the generic version offered by Bachman and Palmer (2010), the decision claim on the CET-4 is articulated in the above box. Articulation of this claim involves listing major types of decisions and qualities that its supportive warrants should be concerned about.

Table 4.1 presents the major types of decisions made on CET-4 scores, the decision-makers and stakeholders to be affected by these decisions. The second row displays decisions made by the NCETC on CET-4 scores, which are publicized to different groups of stakeholders via official statements and the CET-4 Syllabus. The third row is related to decisions made at institutional levels, which are a major source of contention on the CET-4. Related backing will be collected from questionnaires and interviews. The employment decision within the social dimension is listed at the bottom but will not be further explored given that the study is mainly conducted within the instructional setting.

Table 4.1 Major types of decisions made on CET-4 scores

Multiple decisions	**Stakeholders to be affected by the decisions**	**Corresponding decision makers**
Set different CET-4 cut-off scores ● 550 for taking CET-4 SET ● 425 for taking CET-6 ● 220 for issuing the Score Report Form	students teachers the University Academic Affairs Office	the NCETC
Institutional decisions	to be investigated	to be investigated
CET-4 scores as a prerequisite for employment	university graduates	employers

Warrants in support of the decision claim involve two qualities: value sensitivity and equitability. In order to help bring about the intended consequences, test users need to take actions, which involve making decisions. These decisions are expected to bring about some specific beneficial consequences to particular individuals or groups of individuals, and perhaps to an educational program and society (Bachman & Palmer, 2010). Thus, decision makers need to take into account the context of educational and societal value systems, as well as legal requirements and regulations in their decision-making processes. In terms of equitability, decisions that are made should not favor any particular group of test takers. Equitability should be ensured in decision-making procedures, particularly criteria of setting cut scores. Equitability should also be manifested in providing equal opportunities for all test takers to learn and acquire the ability to be assessed. Adapted from the generic versions provided by Bachman and Palmer (2010), the following warrants are articulated in line with the two qualities.

A. Warrants about the value-sensitivity of the decisions that are made:

Warrant A1: Decisions made on CET-4 scores take into account the existing educational and societal values against the background of Chinese testing setting.

Warrant A2: Decisions made at the institutional levels take into account the legal documents, relevant university regulations and common practices in the university community in Xi’an.

B. Warrants about the equitability of decisions that are made:

Warrant B1: The same cut scores are used in making decisions and no other considerations are taken into account.

Warrant B2: Test takers, EFL teachers and other stakeholders within the university community are fully informed about how the decisions are made and whether decisions are actually made in the way described to them.

Warrant B3: Test takers have equal opportunity to learn or acquire the ability to be assessed.

4.4.4 Claim on consequences

Claim 3: The consequences of using the CET-4 and of the multiple decisions that are made are beneficial to stakeholders.

Based on the generic version offered by Bachman and Palmer (2010), the claim on consequences of the CET-4 is articulated in the above box in response to RQ3. The claim of consequences should entail two parts: one pertaining to the intended consequences, and the second to the stakeholders who will be affected by these consequences. I will first discuss these two concerns. Identifying stakeholders at various levels can be helpful to articulate the warrants in support of claim 3. In light of the large-scale and high-stakes nature of the CET-4, the stakeholders directly and immediately affected are College English teachers, university students, and the University Academic Affairs Office. Their teaching and learning practices, the curriculum revision and program development and so forth may be influenced by use of the CET-4 and by decisions that are made on it. Those indirectly and less affected include policy makers, employers, parents and even the public. For example, policy makers in a given educational departments take into account the feedback from stakeholders and statistical analyses from CET-4 scores to make decisions on educational policies or reforms. Employers set CET-4 scores as one of selective criteria for their recruitment. Parents may be burdened if their children are denied of the Bachelor's degree only due to their children's CET-4 scores below the cut-off line. The public may hold positive or negative perceptions towards the assessment.

Since different groups of stakeholders may be influenced to various degrees by an assessment, it is advisable to take into consideration the intended consequences of the CET-4 to decide which stakeholders should be included in the present study. The CET-4 is intended to provide an accurate measure of overall English proficiency of college and university students and to promote effective teaching and learning practices of English as a foreign language at the tertiary level in China. Thus, the present study will target the primary stakeholders at college, namely, English teachers, students, and administrators in the University Academic Affairs Office who are responsible for making institutional policies and program decisions related to the CET-4. In addition, The NCETC will be included as well for its responsibility for test development and setting different cut-off lines. Given the limited resources and practical constrains, stakeholders within the social dimension will not be investigated.

Having determined the targeted stakeholders in this study, we now return to RQ3. RQ3 explores the consequences of the CET-4 on English learning and teaching, namely washback. The present study is more concerned with applying an AUA to link validity issues to washback phenomenon than examining washback per se. Given its complex dimension, investigation of washback in this study will be explored from stakeholders' perceptions of the test and its washback (RQ3.1), while RQ 3.2 attempts to link washback to validity issues.

RQ 3.1 How do stakeholders perceive the CET-4 and its washback?
RQ 3.2 How do students' perceptions affect their test performances?

Articulations of warrants or rebuttals should always be related to research questions and take into consideration the specific CET-4 context. Bachman and Palmer (2010) list five warrants about the beneficence of consequences of using the assessment, involving three aspects: groups of stakeholders, assessment reports, and washback. They also specify concerns about the consequences of decisions that are made. Thus, warrants in support of the consequence claim are articulated as follows:

Warrant 1: The consequences of using the CET-4 are beneficial to immediate stakeholders including students, teachers, as well as the universities.

Warrant 2: The CET-4 Score Report Forms are treated confidentially, presented in clear and understandable ways, and released to test takers and the University Academic Affairs Office in time for them to be used for making decisions.

Warrant 3: The consequences of using the CET-4 and of the decisions made on it help promote desirable instructional practices and effective learning in College English instructional settings.

Warrant 1 is articulated in line with RQ3.1. Backing will be collected from questionnaires and interviews to explore perceptions of administrators in the University Academic Affairs Office, teachers and students to the CET-4. Warrant 2 is particularly applicable to the CET-4 because adopting the Score Report Form to replace the traditional certificate is one of the major reforms conducted by the NCETC. Warrant 3, pertaining to washback, entails investigation into RQ3.2. For example, students' perceptions of the test will affect their learning behaviors, test preparations, which may in turn influence their test performances. Warrant 3 involves all the immediate stakeholders mentioned earlier and covers consequences of test reforms, institutional decisions, aspects of classroom activities, learning behaviors, test preparations, and so on.

While reviewing the evolution of the CET-4 in Chapter 2, it can be noticed that it is the negative washback of the CET-4 that has induced the strongest criticisms (e.g., Cai, 2005, 2006; Gu, 2003, 2007; Han, Dai &Yang, 2004; Jin, 2008; Liu& Dai, 2004). These criticisms identified Warrant 3 as open to challenge, so a concern to the potential rebuttal from other stakeholders is listed below. However, I will let the empirical evidence to be generated in this study to speak as either backing or rebuttal backing.

Rebuttal: The use of the CET-4 generated unintended or negative consequences (such as the phenomenon of "teaching and learning to the test", narrow of curriculum, anxiety).

4.5 Summary

This chapter first linked the three research questions to the claims in the AUA for the CET-4. Then TOEFL iBT, IETLS, and the NMET were discussed to demonstrate qualities that large-scale and high-stakes tests are supposed to possess, which reinforced the legitimacy of drawing the present study on the AUA for evaluation and justification. Following this, I explained the reason why this study mainly focused on three claims pertaining to interpretations, decisions and consequences of the CET-4. Finally, the specific warrants in accordance with sub-questions were articulated and one threatening rebuttal was identified. As Figure 4.2 shows, the framework for the present study was eventually established in this chapter. Justifying test use determines adoption of a bottom-up approach. The inferential links of the AUA for the CET-4 thus start from the claim on interpretation of CET-4 scores, to score- based decisions and to consequences of using the CET-4 and of decisions that are made based on it. The claim on consistency, signified by the dotted box, is not investigated due to the limited access to test scores.

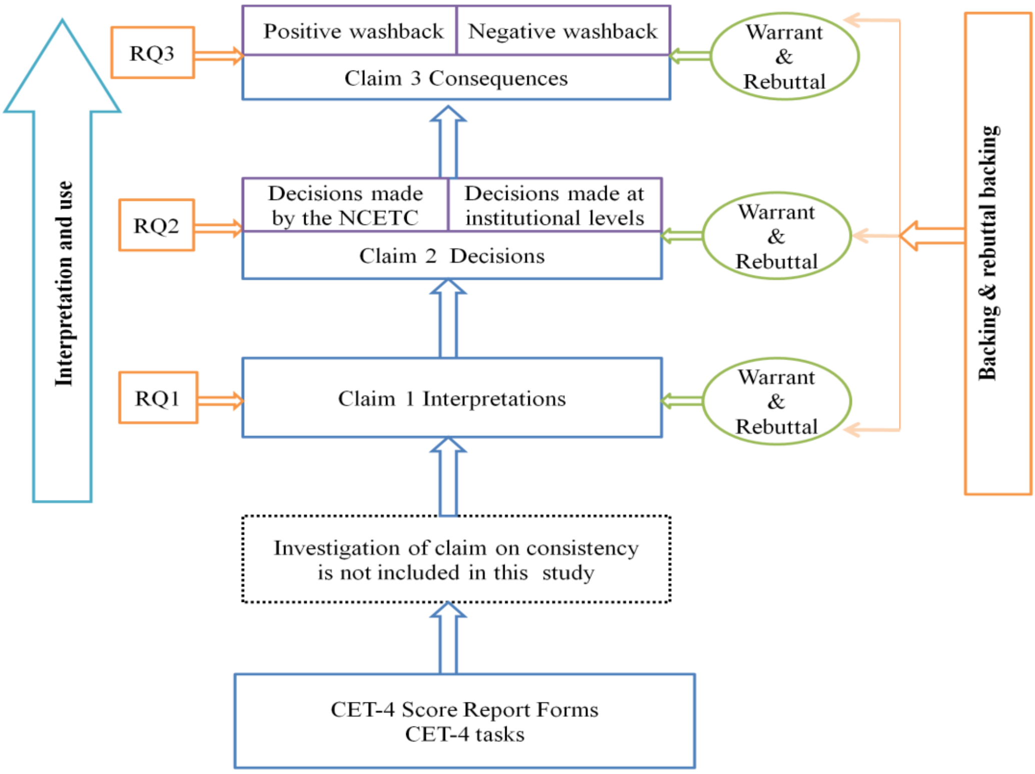

Figure 4. 2 Inferential links in the AUA for the CET-4

Bachman and Palmer (2010) stress that the process of assessment justification consists of two steps: articulation of an AUA and collection of backing to support these statements. Therefore, the next step is to decide on backing sources and the collection methods. Before I introduce the research design of this study, the next chapter will discuss a preliminary study in phase I, in which major instruments like questionnaires and statistical analysis for test data were piloted at this stage. Methods to collect backing or evidence for the main study are to be elaborated on in Chapter 6.

CHAPTER 5

THE PRELIMINARY STUDY

5.1 Introduction

This chapter reports on findings from the preliminary study in relation to construct of the CET-4. The major purposes of the preliminary study were to: 1) seek answers to RQ1.1: to what extent does the CET-4 measure the construct to be assessed, 2) collect part of backing in support of the corresponding warrants 1 and 2 under the interpretation claim of the AUA framework (see section 4.4.2), 3) trial methods and instruments of questionnaire surveys for data collection in the main study. To be more specific, section 5.2 presents findings from a comparative study between the pre- and post-2006 CET-4. Section 5.3 describes a needs analysis, exploring whether the CET-4 can reflect test takers' target language use domain and meet students' demands for English proficiency. Section 5.4 discusses the piloted student and teacher questionnaires

5.2 Comparative study on the two versions of the CET-4

5.2.1 Motivations of the comparative study

This comparative study aimed to investigate, statistically speaking, to what extent the current CET-4 could serve as an improved measure of students' overall English proficiency compared with the CET-4 before 2006. In this chapter, the two versions of the test are addressed as the old CET-4 (pre-2006) and the new CET-4 (post-2006) respectively, since the public more commonly addresses them as the old and the new ones.

This comparative study was motivated by several factors. First, according to Fulcher and Davidson (2009), sometimes new task types are incorporated into an existing test with the purpose to make the test more construct representative, suitable for its intended purposes, and congruent with evolving standards. One of the underlying reasons for the 2006 CET-4 reform is to meet new standards stipulated by the CECR and keep pace with the increasing demands for students' overall English proficiency, especially their listening and speaking abilities. When an upgraded test is designed to replace the old one, it is necessary to provide evidence to justify the necessity and reasonableness of the test reform and convince stakeholders of the improved test quality. This provides a theoretical rational for this comparative study. Second, Chapter 7 examines content validity of the current CET-4, in which a diachronic approach was adopted to make a comparison between the new and the old CET-4 in the aspects of readability and genres (see sections 7.3.1.2 & 7.3.1.3). Echoing the qualitative analyses on test content, quantitative analyses were conducted to examine the internal structure of the two versions of the CET-4. Finally, construct validity has been regarded as a unitary and overarching concept subsuming various aspects of validation (Hughes, 2003). Among the multiple types of evidence, concurrent validity can provide one source of validity evidence. The same group of students took both versions of the test roughly at the same time, which provided a prerequisite and feasibility for this comparability study. With the above reasons, a comparability study between the old and the new versions of the CET-4 was conducted in order to collect evidence related to RQ 1.1 and its corresponding warrants.

5.2.2 Methodology

5.2.2.1 Participants

The test takers were sophomores from six intact classes of a comprehensive university in Xi'an, who just took the CET-4 in December 2008 after finishing one and half a years' English study. They had a variety of English proficiency levels, and academic backgrounds, which included history, biology, physics, tourism management, environmental projects and archeology. One hundred and ninety-two students were recruited to take both versions of the CET-4, of which 51% of students majored in Humanities and Arts, and 49% in Science and Engineering.

5.2.2.2 Instruments

Two test papers were used for test administrations in the comparative study. One was the official test paper of the CET-4 administered on 20 December 2008, which was developed by the NCETC. The other was the university’s final English exam paper in the format of the old CET-4, developed by an experienced EFL teacher and moderated by two CET test administrators, who were also experienced EFL teachers. The reason for choosing the final exam as the criterion measure was two-fold. First, it followed the design of the old CET-4. Second, students would take the final exam seriously, since its credits would be recorded in their academic performance files and failure in the exam meant retaking both the English course and the exam in the next semester. The reason why the study did not choose one authentic pre-2006 test paper to administer a test was that students may have done all the test papers. Their familiarity with the test paper and motivations for taking a test of old-format CET-4 just for a research purpose would undermine their true performances.

5.2.2.3 Data collection

All the 192 students took the official administration of the CET-4 on 20 December 2008, and their final exam in the format of the old CET-4 a week later. In terms of the final exam, the objective part occupying 85% of the total weight were machine scored, and the Writing section taking up 15%, were rated by two teachers following the criteria issued by the NCETC. Both were awarded as excellent official CET-4 essay raters. The final score for each essay was the average of the two raters’ scores. In addition, any essay with score discrepancy beyond three points was rerated and a final score was given after raters’ discussion to ensure reliability. The official CET-4 scores were released on an authorized website in early March of 2009 and the Score Report Forms were issued to the university in the late. With the agreement from both students and the University Academic Affairs Office, 192 pairs of the old and new CET-4 scores had been obtained by the end of April.

5.2.2.4 Data analysis

Two sets of test data were analyzed by SPSS 16.0. The statistical analyses for test scores included basic descriptive analysis, correlations between the total scores and the subtest scores in both the old and new versions of the test, intercorrelations between the two tests, shared variance, and exploratory factor analyses (EFA).

5.2.3 Results

5.2.3.1 Descriptive analysis

The purpose of the analysis was to make a comparison between the old CET-4 and the new CET-4 to see whether there is any significant variance. First, the basic descriptive analysis was conducted to the old and the new CET-4respectively. Four cases with missing values and outliers were deleted, making 188 sets of test scores as the final valid data points.

As Tables 5.1 and 5.2 show, the total scores of the old CET-4 ranged from 38 to 89 with a standard deviation of 9.40 on a 100 score scale, and the total scores of the new CET-4 ranged from 333 to 626 with a standard deviation of 57.25 on a 710 score scale. Each indicated a satisfactory score distribution. The reliabilities of the old and new versions were .78 and .79.

Table 5.1 Descriptive statistics of the old CET-4

Old CET-4	Min	Max	Mean	SD	Skewness		Kurtosis	
OLC	3.00	18.00	11.34	3.00	-.112	.177	-.337	.353
OR	18.00	40.00	31.86	3.88	-.668	.177	1.145	.353
OVS	2.50	13.00	8.72	2.24	-.321	.177	-.625	.353
OCL	2.50	10.00	5.86	1.52	.021	.177	-.265	.353
OWR	4.00	13.00	8.17	1.95	.356	.177	-.097	.353
OTOT	38.00	89.00	65.92	9.40	-.109	.177	-.220	.353

Note. OLC=old listening comprehension; ORC= old reading comprehension;
OVS= old vocabulary and structure; OCL=old cloze; OWR= old writing; OTOT= old total.

Table 5.2 Descriptive statistics of the new CET-4

New CET-4	Min	Max	Mean	SD	Skewness		Kurtosis	
NLC	108	224	167.24	25.48	-.024	.177	-.675	.353
NRC	99	228	166.12	25.05	-.141	.177	-.584	.353
NCL	29	70	47.54	8.457	-.163	.177	-.290	.353
NWT	58	124	92.15	13.63	-.107	.177	-.374	.353
NTOT	337	626	472.94	57.25	-.104	.177	-.555	.353

Note. NLC= new listening comprehension; NRC= new reading comprehension; NCL=new cloze; NWT= new writing & translation; NTOT= new total.

5.2.3.2 Correlation analysis

Correlation is a common statistical analysis in conducting construct validity research. In order to evaluate the construct validity of the test paper, a Pearson's product-moment correlation was conducted to explore the relationship between each subtest with other subtests and with the total test. According to Alderson et al., (1995, p.184), the correlations among different subtests might be expected to be lower since they all measure something different, while the correlations between each subtest and the whole test might be expected higher since the overall score is taken to be a more general measure of language ability than each individual component score. Although Yang and Weir (1998) suggested in their CET-4 validation study that it should be acceptable for the correlation coefficient between the subtests to be within a range of .30-.70, it should be noted that the coefficient below .40 is regarded as low.

Table 5.3 displays the correlation matrix of the old CET-4. All the correlations were significant. The correlations between the total score and all the five subtest scores were moderately higher (.696-.785), topped with the correlation between the reading and the total (.785). The correlations among the subtests were fairly low in a range of .364 to .544. The lowest one (.364) was between writing and reading components, which may be explained by the fact that the two components examined different receptive and productive skills.

Table 5.3 Correlations of the old CET-4

Subtest	OLC	ORC	OVS	OCL	OWR	OTOT
OLC						
ORC	.422**					
OVS	.380**	.385**				
OCL	.544**	.442**	.499**			
OWR	.513**	.364**	.462**	.412**		
OTOT	.777**	.785**	.696**	.722**	.699**	

Note. OLC=old listening comprehension; ORC= old reading comprehension; OVS= old vocabulary and structure; OCL=old cloze; OWR= old writing; OTOT= old total
**.Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

For the new CET-4, Table 5.4 shows that the correlation between each subtest ranged from .317 to .555. The correlation between listening and reading (.547), and that of listening, and writing and translation (.555) were at the top rank.

Table 5.4 Correlations of the new CET-4

Subtest	NLC	NRC	NCL	NW&T	NTOT
NLC					
NRC	.547**				
NCL	.335**	.343**			
NW&T	.555**	.436**	.317**		
NTOT	.867**	.834**	.521**	.722**	

Note. NLC= new listening comprehension; NRC= new reading comprehension; NCL=new cloze; NWT= new writing & translation; NTOT= new total.
**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

The results could be explained from two aspects. First, listening, reading and writing were all basic skills, so students' performances in these three subtests were consistent. Second, the compound dictation in the new CET-4 listening component tested students' both receptive and productive skills, which were also measured by writing and translation part. In addition, strong and positive correlations were observed between listening and the total score (.867) and between the reading and the total (.834), which showed their importance and strong effect on the total score. It can be seen from Table 5.4 that the correlations between cloze and other subtests were below .40. The correlation between cloze and the total was comparatively lower (.521). Part of the explanation may be that the cloze part had lower validity, and partially because some students gave up the cloze part for insufficient time, which was evidenced by the later questionnaire and interview findings.

Table 5.5 illustrates the intercorrelations between all components of the two tests. The total scores of the two tests were highly correlated (.754), and the correlation between the scores on the listening subtests of both tests was also fairly strong (.701).

Table 5.5 Correlations between the old and the new CET-4

	OLC	ORC	OVS	OCL	OWR	OTOT	**NLC**	**NRC**	**NCL**	**NWT**	**NTOT**
OLC											
ORC	.422**										
OVS	.380**	.385**									
OCL	.544**	.442**	.499**								
OWR	.513**	.364**	.462**	.412**							
OTOT	.777**	.785**	.696**	.722*	.699**						
NLC	.701**	.494**	.466**	.510**	.544**	.733**					
NRC	.446**	.363**	.456**	.416**	.432**	.558**	.547**				
NCL	.253**	.268**	.300**	.432**	.210**	.377**	.335**	.343**			
NWT	.405**	.325**	.452**	.367**	.501**	.536**	.555**	.436**	.317**		
NTOT	.643**	.496**	.557**	.561**	.582**	.754**	.867**	.834**	.521**	.722**	

Note. OLC=old listening comprehension; ORC= old reading comprehension; OVS= old vocabulary and structure; OCL=old cloze; OWR= old writing; OTOT= old total; NLC= new listening comprehension; NRC= new reading comprehension; NCL=new cloze; NWT= new writing & translation; NTOT= new total.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

However, there was a weak correlation between the reading components of the two tests (.363). In comparison, the correlations between the listening and the total(.867) and the reading and the total(.834) in the new CET-4 were higher than their counterparts in the old CET-4 (.777, .785), which could be regarded as convincing evidence that new CET-4 displayed a better internal structure. The statistics generated in this study are in line with the findings of the NCETC. According to Jin (2006), in the study conducted by the NCETC, the released correlations among the subtests of the new CET-4 (trial version) ranged from .549 to .732, much higher than their counterparts in the old version. It indicates that the new CET-4 to some extent better measures students' overall language ability from multiple facets and perspectives.

In order to examine the proportion of overlap between the two sets of test data as a way of determining concurrent validity, the shared variances (R^2) of the correlation coefficients between the reformed components of the two tests were produced based on the correlation coefficients (see Table 5.6).

Table 5.6 Shared variances of the correlation coefficients of the two tests

Test	OLC	ORC	OTOT
NLC	.491		
NRC		.132	
NTOT			.569

Note. OLC=old listening comprehension; ORC= old reading comprehension; OTOT= old total; NLC= new listening comprehension; NRC= new reading comprehension; NTOT= new total.

The value of .491 indicated that there was 49.1% overlap of the variance between the scores on the listening components of the old and the new CET-4. The low value of .132 indicated the two reading subtests measured different skills, since the construct of fast reading was different from that of careful reading. The value of shared variance (.569) between the two sets of total scores indicated that almost 43.1% of the constructs of the two tests was different from each other, suggesting that while the two tests measure the same general areas of skills such as listening and reading, they in fact involved some different aspects of the constructs. The findings are consistent with what Dorans, Moses and Eignor (2011) suggest. When the same examinees take both tests, the direct control over differential examinee ability can be achieved. In other words, since the same test population took both tests, it can be reasonably assumed that the unexplained parts of the variances were not caused by different test taker characteristics but by differences in either test constructs or test methods of the two versions of the CET-4.

5.2.3.3 Exploratory factor analysis

This study adopted exploratory factor analysis rather than confirmatory factory analysis (CFA) to analyze the test data for the following reasons. First, the study did not specify a priori hypothesis on how many factors would be extracted and how many variances the factors would comprise, so this step was exploratory instead of being confirmatory in nature. Second, with limited test data, CFA like structural equation modeling cannot be performed in this study. Thus, in order to know which components in the old and new CET-4 played greater role in contributing to the total score, exploratory factor analysis was conducted. Principal components analysis (PCA) and factor analysis (FA) are two statistical techniques to extract factors. The goal of PCA is to extract maximum variance from the data set with each component, while in FA only the variance that each observed variable shares with other observed variables is available for analysis. In practice, PCA is a better choice to obtain an empirical summary of the data set. FA is used to seek a theoretical solution uncontaminated by unique and error variability (Tabachnick & Fidell, 2001). Rotation of factors is a process to make the solution more interpretable without changing its underlying mathematical properties. There are two general classes of rotation: orthogonal and oblique. The former assumes that variables are independent of each other while the latter supposes that variables may be correlated to each other (Tabachnick & Fidell, 2001). In terms of an assessment, different subtests are assumed to measure different abilities of test takers but they are not completely independent of each other, altogether contributing to an overall construct. Therefore, principal components analysis was adopted as the extraction method, with oblique as the rotation approach because the underlying constructs are assumed to be correlated.

Tables 5.7 and 5.8 display results of KMO and Bartlett’s test of the old and new CET-4. KMO and Bartlett’s test is to check whether the distribution values are adequate for conducting factor analysis. Conventionally, a KMO measure >.9 is marvelous, >.8 is meritorious, >.7 is middling, >.6 is mediocre, >.5 is miserable, and <.5 is unacceptable for factor analysis (George & Mallery, 2007, p.256). In this case, the KMO values of the two versions of the CET-4 were .802 and .743, indicating the acceptability for EFA. Bartlett’s Test of Sphericity is a measure of multivariate normality of a set of distributions. It also tests whether the correlation matrix is an identity matrix because factor analysis would be meaningless with an identity matrix. A significance value at $p < .05$ indicates that these data do not produce an identity matrix (George & Mallery, 2007, p.256). As shown in the two tables, the significance of both tests is .000, indicating that the two sets of test data are approximately multivariate normal and acceptable for factor analysis.

Table 5.7 KMO and Bartlett's Test of the old CET-4

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.802
Bartlett's Test of Sphericity	Approx. Chi-Square	260.195
	df	10
	Sig.	.000

Table 5.8 KMO and Bartlett's Test of the new CET-4

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.743
Bartlett's Test of Sphericity	Approx. Chi-Square	173.568
	df	6
	Sig.	.000

Tables 5.9 and 5.10 show the Eigenvalues of the two sets of test data, which are the proportion of total variance in all the variables accounting for the factor.

Table 5.9 Total variance explained of the old CET-4

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.773	55.468	55.468	2.773	55.468	55.468
2	.659	13.171	68.639			
3	.626	12.511	81.150			
4	.563	11.264	92.413			
5	.379	7.587	100.000			

Table 5.10 Total variance explained of the new CET-4

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.284	57.090	57.090	2.284	57.090	57.090
2	.746	18.638	75.728			
3	.563	14.070	89.798			
4	.408	10.202	100.000			

The ratio of Eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. Only extracted factors with Eigenvalues of 1.0 or higher contribute to the explanation of variances in the variables (Kinnear & Gary, 2008, p.552). The "Initial Eigenvalues" and the "Extraction Sums of Squared Loadings" columns are the same, except that the latter only lists factors which have actually been extracted. From Table 5.9 it can be seen only one factor (Eigenvalues 2.773, higher than 1.0) was extracted from the five components for analysis, which accounted for 55.468% for the variance. Table 5.10 displays similar results that one factor in the new CET-4 was extracted (Eigenvalues 2.284), accounting for 57.09% for the variance.

Findings from factor analysis of the old CET-4 are the same as conclusions reached by Yang and Weir (1998) in their CET-4 validation studies in which they interpreted factor one as the general English ability. For the new CET-4, factor one can be interpreted as the overall English proficiency, especially the abilities involved in listening. Since only one factor was extracted from each test respectively, which was not sufficient to reveal variance between them, the factor loadings of different components in both versions of the CET-4 were further examined.

The "Component Matrix", as the central output of the factor analysis, displays the factor loadings. The factor loadings of components in the old CET-4 were between .787 and .690 (see Table 5.11). The Cloze had the highest loadings (.787), which can be explained by the fact that Cloze is regarded as a component requires comprehensive skills, especially the grammatical and lexical knowledge.

Table 5.11 Component Matrix of the old CET-4

	Component
Cloze	.787
Listening Comprehension	.775
Writing	.739
Vocabulary & Structure	.729
Reading Comprehension	.690

Note. Extraction Method: Principal Component Analysis

The factor loadings of components in the new CET-4 were between .830 and .617 (see Table 5.12). The Listening has the highest factor loading (.830), followed by the loadings of the Reading component (.782). The findings corresponded to the College English teaching objective of developing students' ability to use English in an all-round way, especially in listening. It was also consonant with revisions made in the 2006 CET-4 reform to increase weight on measuring the listening ability. From this perspective, the new CET-4 demonstrates better internal structure of construct.

Table 5.12 Component matrix of the new CET-4

	Component
Listening Comprehension	.830
Reading Comprehension	.782
Writing& Translation	.776
Cloze	.617

Note. Extraction Method: Principal Component Analysis

5.3 Needs analysis

In response to RQ1.1 and RQ 1.2, and the warrants under the interpretation claim (see section 4.4.2), a needs analysis was conducted to examine the constructs to be assessed by the CET-4. Needs analysis aims at collecting and analyzing information necessary to meet students' language learning needs within the context of a particular institution involved in the learning or teaching situation (Brown, 1996, 2001). The primary purpose of language testing is to make inferences about test takers' ability to use language in a TLU domain. Thus, test designing should incorporate tasks with test features corresponding to those of target language use tasks (Bachman & Palmer, 1996, 2010).

Therefore, in order to examine to what extent the coverage and the relevance of test tasks are aligned with the intended purpose of the reformed CET-4 and are representative of TLU tasks, a small group of survey items were designed to explore university students' needs for English communication in real-life situations (see Appendix C). Moreover, the investigation into the authenticity of the CET-4 by the needs analysis is expected to evidence the content validity of the CET-4.

The needs analysis questionnaire was distributed to 232 students, including 164 test takers from the above-mentioned old and new versions of the CET-4 administration and 68 juniors who took the CET-4 one year or half a year ago. Altogether 229 questionnaires were collected. The response rate was 98.7%. Ten questionnaires with more than half of the unanswered items were discarded. Finally, 219 student questionnaires were kept as valid. Among them 46% of respondents were male students, 54% female, and 39% were from the Humanities and Arts majors and 61% from the Engineering and Science. The overall reliability of the needs analysis items was .871.

When students were asked about the circumstances in which they would use English in their daily or future work-related interactions, watching TV programs and films (Mean=3.83), taking English test (Mean=3.45), and attending lectures delivered in English (Mean=3.31) ranked the top three based on a 5-point Likert scale of frequency. The top three text types that students would read were textbooks (Mean=3.92), test papers (Mean=3.75), newspapers and magazines (Mean=3.09). The text types that they may write in English were listed in order of descending means: thesis abstract, e-mail, diary, and memo. Their means were all below 3 points, indicating students thought they would only occasionally write them.

Exploratory factor analysis was conducted on the needs analysis items. First, principal component analysis was adopted as factor extraction method. The KMO value (.845) and the significance value (.000) in Bartlett’s test of sphericity showed the appropriateness of this statistical analysis (see Table 5.13).

Table 5.13 KMO and Bartlett's Test[a] of needs analysis

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.845
Bartlett's Test of Sphericity	Approx. Chi-Square	190.643
	df	210
	Sig.	.000

Table 5.14 shows five factors (whose Eigenvalues >1) were extracted from 21 items, totally accounting for 62.44 % of variance of the variables.

Table 5.14 Total variance explained of the needs analysis

Component	Initial Eigenvalues[a]			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.631	31.574	31.574	6.631	31.574	31.574
2	2.297	10.939	42.513	2.297	10.939	42.513
3	1.591	7.578	50.091	1.591	7.578	50.091
4	1.384	6.592	56.683	1.384	6.592	56.683
5	1.210	5.762	62.444	1.210	5.762	62.444
6	.949	4.521	66.965			
7	.852	4.059	71.025			
8	.700	3.332	74.357			
9	.667	3.178	77.535			
10	.620	2.952	80.487			
11	.605	2.880	83.366			
12	.553	2.632	85.999			
13	.499	2.377	88.375			
14	.425	2.023	90.399			
15	.386	1.840	92.239			
16	.359	1.711	93.950			
17	.317	1.510	95.459			
18	.312	1.487	96.946			
19	.277	1.321	98.267			
20	.202	.962	99.229			
21	.162	.771	100.000			

In terms of rotation, oblique rotation was performed first, but the results turned out to be hard to interpret. In particular, the component correlation matrix indicated poor correlations among the extracted factors. Then orthogonal rotation was conducted and varimax with Kaiser Normalization was adopted as the final rotation method. For one thing, the results were more meaning and interpretable. For another, the needs analysis was aimed at identifying different scenarios in which students are required to use English, so factors could be assumed not to correlate with each other. Table 5.15 presents rotated component matrix of the needs analysis. Factor 5 was deleted because it only had one item. Thus, a four-factor solution was generated.

Table 5.15 Rotated component matrix of the needs analysis

Items	F1	F2	F3	F4	F5
9	**.788**	.147	.348	-.034	-.019
8	**.772**	.119	.269	-.018	-.052
5	**.760**	.290	.012	.036	.009
4	**.729**	.110	.087	.126	-.012
2	**.657**	.077	.023	-.186	.222
17	**.494**	.447	.281	-.012	-.149
12	.160	**.855**	.087	.011	-.118
13	.227	**.810**	.196	.049	.013
19	.138	**.796**	.089	.011	.201
20	.156	**.573**	.475	-.051	.258
21	.193	**.504**	.444	-.066	.253
16	.131	**.464**	.418	.214	-.178
15	.049	.112	**.810**	.038	.101
14	.208	.183	**.700**	.113	-.048
7	.311	.088	**.535**	.071	-.295
18	.033	.296	**.526**	.010	.485
6	.457	.202	**.524**	-.029	-.101
11	-.063	-.126	.108	**.828**	.085
3	.113	.129	-.027	**.739**	-.030
10	-.289	.083	.152	**.533**	.433
1	.085	.012	-.102	.112	**.806**

Note. Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization

As Table 5.15 displays, factor 1, accounting for 31.57% of variance of all the observed variables, was extracted from six items (9, 8, 5, 4, 2, 17). Item 17 was dropped for poor theoretical fit and comparatively lower factor loadings (.50). It is evident that factor 1 is concerned with both formal and informal interactions such as travelling abroad, talking with foreigners, academic exchanges, interviews, etc in which effective communication skills in English especially speaking and listening abilities are highly demanded.

Factor 2 accounting for 10.94% of the variance of all the variables was extracted from six items (12, 13, 19, 20, 21, 16), which can be interpreted as the ability of reading and writing professionally in academic or work-related domains such as reading major-related literature, writing thesis abstracts.

Factor 3 accounting for 7.58% of the variance of all the variables was mainly extracted from five items (15, 14, 7, 18, 6), which is related to reading and writing in English for interests and entertainment. The text types cover newspapers, magazines, novels, English websites, TV programs and movies, and diary.

Factor 4 accounting for 6.59% of the variance of all the variables was extracted from three items (11, 3, 10), which involves test-taking domains and test taking strategies.

To sum up, factor 1 occupying the largest variance indicates a strong and urgent need to cultivate students' listening and speaking abilities, which is in line with the teaching objectives stated by the CECR. In addition, according to the CET-4 Syllabus, the test content should mainly cover domains of daily conversations, lectures, radio and TV programs, newspapers, magazines, books and academic journals. The text types students are likely to hear, read and write in English evidence the selective criteria of CET-4 content materials. The above needs analysis suggests that CET-4 test developers have attempted to improve its authenticity and to make the test tasks more similar to the target language use domain.

5.4 Piloted Questionnaires

Both student and teacher questionnaires were piloted on a small scale two days after administration of the official CET-4 in December. 2008. The student questionnaire explored students' attitudes mainly in the following aspects: their general perceptions of CET-4 and English learning, evaluations on the content and quality of CET-4, attitudes toward test preparation and the future reforms, factors influencing test difficulty and test takers' own problems in doing the subtests, etc. The teacher questionnaire consisted of similar themes.

The item types include 5-point Likert scales, selective responses, order ranking, and open-ended questions. The student questionnaire was in Chinese for the sake of time efficiency and understanding accuracy, while the teacher questionnaire was delivered in English. It took respondents ten to fifteen minutes to complete a questionnaire.

The student questionnaires were administered along with the needs analysis survey to the same group of respondents. About 229 student questionnaires were returned and 219 questionnaires were regarded as valid. Ten teachers were invited to complete the teacher questionnaire. The overall reliability of the student questionnaire is 0.826, and the teacher questionnaire was 0.874 (Cronbach's Alpha).

The major purpose of the piloted questionnaires is to identify ambiguous, confusing items to respondents. In addition, a pilot study also has functions, principally to increase the reliability, validity and practicability of the questionnaire (Oppenheim, 2000; Morrison, 1993; Wilson & McLean, 1994). Hence, the detailed survey findings can be left to the main study report. The following part will briefly talk about major modifications to be made and variables to be included related to validity of revised components and washback of the CET-4.

First, according to the teacher respondents, since the present study is targeted at the revised listening and reading components, some of the items related to content and design of cloze and writing parts were advised to be cancelled. In addition, a few items were presented in different wording but actually elicited overlapping responses. The redundant items were discarded.

Second, some item types were changed in order to elicit revealing findings for the main study. For example, when students were asked about the motivations for learning English and taking the CET-4, the items were designed in selective responses. It turned out that students chose more than one and even all the options, which indicated they were multi-motivated by different factors. Given that the selective response item type failed to reveal the motivating degree of all the factors, in the main study students will be asked to rate their learning and test taking motivations based on a 5- point Likert scale of agreement. Another change was made in the open-ended question type. It advantage is to enable respondents to share whatever opinions and concerns they hold so that unexpected information may emerge. In other words, some aspects that are neglected for exploration, based on respondents' output, may be identified. The disadvantage lies in its time-consuming analysis. Given that most of findings from open-ended questions evidenced what has been known from selective responses. In the main study, this item type will not be included.

Finally, the most important contribution of the piloted questionnaire is that some variables were identified for further exploration. To name a few, the survey findings revealed that the majority of students believed test preparations could improve their CET-4 performance, and they admitted that the normal teaching would be replaced by test preparation courses upon approaching the CET-4 administration. It is therefore important to add items investigating teaching activities in test preparation courses and students' own preparation methods out of class. In the piloted student questionnaire, students rated the factors influencing their listening and reading performance. However, how they handle these difficulties in the test also deserves investigating. Hence, items exploring students' test taking strategies will be added. In addition to the above mentioned, items related to test influences will be expanded.

5.5 Summary

This chapter has made a statistical comparative study between the old (pre-2006) and the new (post-2006) versions of the CET-4. A series of statistical analyses were conducted with two sets of scores from the two versions of the CET-4. Detailed statistical results have been described and discussed above (see section summary, p.126). The statistical analysis of the old CET-4 generated similar findings with what Yang and Weir (1998) discovered in their three-year validation study. The statistics from the new CET-4 administered in December 2008 were consistent with those revealed by the NCETC in their research on the trialed version of the new CET-4 held in June 2006. Both studies find that the correlations of the subtests with the total of the new CET-4 are higher than those of the old CET-4 (Jin 2006), indicating that the current CET-4 demonstrates a better internal test structure. In addition, the intercorrelations and the shared variances between the two tests also suggest that the current CET-4 is a reformed and improved version from the statistical comparative perspective. Above all, it can be concluded that the two versions of the CET-4 overall measure the same general English proficiency including general areas of listening and reading skills, but involve different aspects of listening and reading constructs.

Needs analysis has revealed the gap between students' actual English levels and their expected command of English in the future life and work related situations. It in turn demonstrates that the 2006 CET-4 reform has taken into account students' needs for English use in real life situations. The adjusted orientations to improving students' listening and speaking abilities are in line with the social demands for graduates' communicative English proficiency.

This chapter concluded with descriptions of piloted questionnaires. Ambiguous expressions and confusing wording were fine toned. More variables will be included, mainly related to students' test-taking strategies, CET-4 preparation activities, and the affective influences of the test on stakeholders.

Given the above findings and the purposes achieved, the main study will only target at the current CET-4. More advanced statistical analysis will be conducted with a larger pool of test data to explore construct of the revised listening and reading components. Questionnaires and interviews will be distributed on a large-scale to examine test uses and test consequences for triangulation and in-depth investigation of research questions, corresponding claims and warrants, which will be elaborated on in the next chapter.

CHAPTER 6

RESEARCH METHODOLOGY FOR THE MAIN STUDY

6.1 Introduction

In order to answer the research questions and collect backing evidence to support the warrants and claims, both qualitative and quantitative methods are adopted in the present study. This chapter starts with a general account of the research design, justifies the rationality of a mixed-method research design, and specifies the backing collection methods in line with each claim. It also presents a detailed description of participants, research instruments, data collection procedures, and preliminary data analysis.

6.2 Overall research design

Research design can be defined as the overall plan for a piece of research, situating the researcher in the empirical world, and connecting the research questions to data (Denzin & Lincoln, 1994; Punch, 2009). A research design includes four main ideas: strategy to establish a logical rationale behind the design for answering the research questions, the conceptual framework, the question of who or what will be studied, and the tools to be used for collecting and analyzing empirical materials (Punch, 2009). This section will discuss strengths and weaknesses of quantitative and qualitative approaches to offer a rationale for the adoption of a mixed-method design in this study. It follows to specify the major methods for backing collection in accordance with each research question and its corresponding claim.

6.2.1 A mixed-method research

Since the middle of the 20th century, there have been longstanding paradigm debates in education and psychology between advocates of positivistic, quantitative research methodology and advocates of naturalistic, qualitative research methodology (Lincoln & Guba, 1985; Lynch, 1996). A simplified distinction between the nature of quantitative and qualitative data in social research is essentially the distinction between numerical and nonnumerical data (Babbie, 2007; Punch, 2009). A full distinction can be broadened to include ways of conceptualizing the reality being studied and methods.

First, quantitative research tends to conceptualize reality in terms of variables, then measure these variables and study relationships between them (Punch, 2009, p.211). Qualitative research is multi-method in focus, emphasizing observing, describing, interpreting and understanding a specific program, practice, setting or how events take place in the real world rather than in a controlled, laboratory setting (Denzin & Lincoln, 1994; Lynch, 1996; Mertens, 2003). Second, quantitative data are usually collected through more structured procedures, involving a large sample to satisfy statistical requirements. By contrast, qualitative research procedures for data collection are more flexible and dynamic, using techniques such as participant observation, in-depth interviews, and document analysis for gathering and recording data from a variety of sources. However, the qualitative approach usually involves a small sample because an in-depth study requires enormous time and energy (Guba, 1978; Guba & Lincoln, 1989; Lynch, 1996; Wen, 2001). Third, quantitative findings are believed to be more powerful and convincing, more precise and scientific, but tend to be oversimplified and too abstract. In comparison, qualitative outcomes are expected to reveal more in-depth and complex phenomena, but on the other hand they tend to be time consuming and lack of generalizability and representativeness (Cohen & Manion, 2000; Johnson & Onwuegbuzie, 2004; Keeves & Sowden, 1992; Wen, 2001).

Despite the quantitative-qualitative debates, it has been widely assumed that both have their strengths and weaknesses and they should be combined as appropriate. The combination of quantitative and qualitative methods is now increasingly common, and is known as mixed-method research.

A mixed-method research is defined as the third research paradigm, mixing or combining quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study (Brewer & Hunter, 1989; Johnson & Onwuegbuzie, 2004; Punch, 2009; Tashakkori & Teddlie, 1998, 2003; Onwuegbuzie & Leech, 2004). As noted by Johnson and Onwuegbuzie (2004), given that both quantitative and qualitative research is important and useful, a mixed-method research is not intended to replace one for the other but rather to draw from the strengths and minimize the weaknesses of both. Based on the existing literature, the study proposes a mixed-method triangulation design to complement quantitative and qualitative data. Triangulation, involves gathering, reconciling, and explaining of data from multiple methods and multiple sources to avoid the bias inherent in any one particular source or method and to support the strength of interpretations and conclusions (Denzin, 1970; Lynch, 1996; Mertens, 2003). A combination of data sources is likely to be necessary in most evaluations because often no one source can describe adequately such a diversity of features as is found in educational settings, and because of the need for corroboration of findings by using data from these different sources, collected by different methods and by different people.

6.2.2 Methods for collecting backing evidence

The process of assessment justification consists of two sets of interrelated activities: articulation of an AUA and collection of backing evidence. The claims, warrants and rebuttals specific to the CET-4 have been articulated in Chapter 4, in support of the links from assessment performance to consequences of test use. An AUA also serves as a framework for identifying kinds of evidence needed to support these statements, and for collecting backing evidence. One of its merits lies in guiding the study to focus on validity issues that are relevant to a particular test and its purpose so that time and resources can be effectively allocated to collect evidence only bearing upon specific warrants and rebuttals. According to Bachman and Palmer (2010), a wide range of sources and traditional methods for data collection are still available for either backing or rebuttal evidence in an AUA, including documents, regulations, legal requirements, theory, prior research or experience, the procedures in developing and administering the assessment, and procedures in scoring the test takers' responses.

To better answer the research questions and to improve the validity of the study backing will be collected from multiple sources to achieve data and instrument triangulation. The quantitative approach will be applied in the large volume of test data, and questionnaire involving a large number of participants in order to achieve better generalizability of findings. The qualitative approach of interviews, test content analysis, and document analysis will be used to further explain or elaborate on the quantitative results and to increase in-depth and complexity of outcomes. The following part will address specific research methods to be employed by linking them to each research question and its corresponding claim.

For RQ1 and the interpretation claim: Statistical analyses of CET-4 scores and test content analysis will be applied to complement each other to probe into the construct of the CET-4, content representativeness and relevance of test tasks.

For RQ2 and the decision claim: Document analysis, questionnaires and interviews serve as the major instruments to review the multiple decisions made based on CET-4 scores and to explore the factors underlying the decision-making process.

For RQ3 and the consequence claim: Questionnaires and interviews will be mainly employed to investigate in what way and to what extent the CET-4 and the decisions that are made on it affect English teaching and learning practices.

Students' questionnaires will be linked to their test performances to discuss validity along with washback issues.

One point needs to be stressed here. Since an AUA is to provide an overall evaluation of the intended interpretations and uses of scores by generating a coherent analysis of all the evidence available for warrants and rebuttals, each claim with its subordinating warrants and rebuttals may require more than one method to provide the relevant evidence support. This is also a decisive determinant for adoption of a mixed method design. Figure 6.1 illustrates the links between research methods and an AUA framework for the CET-4. The three rectangles on the right list specific methods to collect backing in support of the warrant and claims.

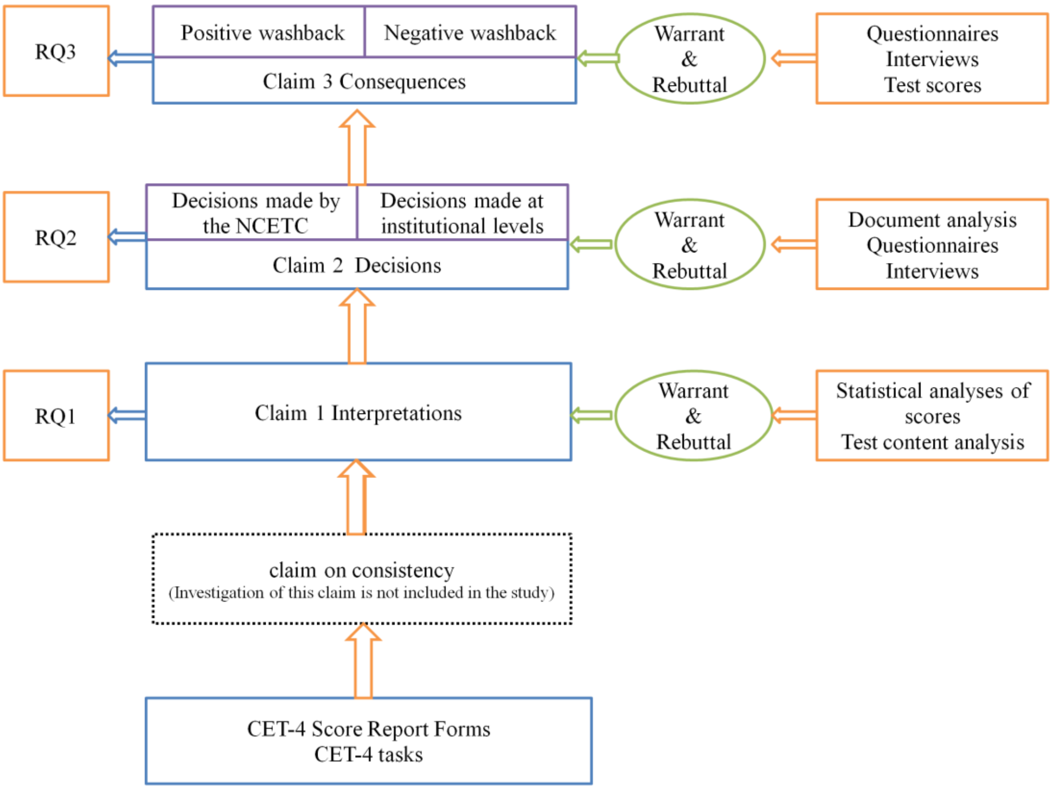

Figure 6.1 Links between research methods and research questions in the AUA for the CET-4

6.2.3 Backing collection site

After the research design was determined, a decision needed to be made on the data collection site prior to sampling participants. With different geographic locations, universities in China vary greatly in terms of the government funds, educational budgets, students' enrollment, teachers' qualifications, teaching and learning resources, and so on. It would be desirable to sample subjects from universities in different parts of China for generalizability and representativeness. However, given limited time and resources, the present study narrowed down its scope and selected Xi'an, capital of Shaanxi Province as the data collection site for the following reasons. Since Xi'an is my hometown and where I used to work, more help and cooperation could be obtained from my former classmates and colleagues who are College English teachers of different universities in Xi'an. Their cooperation would provide easy access to data needed. Second, Xi'an, as one of the important bases of China's higher learning institutions, boasts the largest number of universities among the cities of Northwest of China. These universities range from nationwide prominent to ordinary levels, admitting students from both the northwest regions and other parts of China. Another consideration is that Xi'an is one of the CET essay marking centers, where raters from different universities in Shaanxi Province would be assembled for essay marking. It is convenient to recruit teachers from different universities for the questionnaire survey to enhance the representativeness of findings. Although the study aims at investigating the use of the CET-4 within university community of Xi'an, it is expected to be representative of and generalizable to the contexts in the Northwest part of China.

Eventually four universities (U1, U2, U3, U4) in Xi'an were sampled for the present study. U1 is a comprehensive, multi-disciplinary university under the administration of Shaanxi Province. The other three are science–oriented universities directly under the administration of Ministry of Education. Among them U4 is not only a top-tier university with nationwide fame, but also one of the six universities where the CET-4 norm group was established.

The above section has presented a clear picture of the overall research design and a rationale behind it. The following sections will discuss each data-collection method in detail, including participants, instruments, data collection and analysis procedures.

6.3 CET-4 test data

6.3.1 Participants

Participants for the test score analysis were students of 2008 Cohort in a comprehensive university (U1) in Xi'an. They took the CET-4 after finishing one and half a year's College English learning. The targeted CET-4 in the main study was administered on 19 December, 2009. All the sophomores in this university were required to take the test except students from the School of Arts, who were exempted from taking it for the following reasons. The MoE issued a nationwide policy allowing each province to set a lower NMET cut score as a reference for their university admission. Therefore, students of the School of Arts are usually admitted with a comparatively lower English proficiency. Second, the university sets different policies and English curriculum design for these students. The textbooks they use and the examinations they take are all different from those of other majors. Their graduation degrees are not associated with their CET-4 performances. They can make their own decisions on whether to take the CET-4 or when to take it.

6.3.2 Backing collection

In March 2010, test takers' scores were first released on the official website authorized by the NCETC. Students could log on the website with their CET-4 test taker Identity Numbers and names. The Score Report Forms were distributed to universities in April. With the permission and cooperation of the Academic Affairs Office in U1, the CET-4 score file package of students of 2008 Cohort was obtained for analysis.

A total number of 2948 students in U1 sat for the CET-4 in December 2009. Since students from the School of Arts are not required but encouraged to take the CET-4, only a small number of students volunteered to take the test. Considering this group of students is not homogeneous with the whole test population in U1 due to the different variables mentioned previously, their test cases were deleted and 2777 cases remained. Next, any test score case with either total score or subtest score being zero was deleted from the whole test data pool. The final valid data point is 2692 test score cases.

6.3.3 Analysis of backing from test data

Students' test scores were input into SPSS 16.0 for statistical analysis. Each test case includes a test taker's composite score and the profile scores to the four components. A total number of 2692 data points underwent the following statistical analyses.

Step 1: Descriptive analysis. Basic descriptive statistical analysis was conducted to check the normality and the distribution of scores, as well as assumptions that underline any inferential statistics.

Step 2: Correlation. As a common statistical analysis in conducting construct validity research, correlations at subtest level were conducted to examine the strength of the relationships between each subtest with other subtests and with the total test. In addition, Cronbach's alpha was calculated to inspect how reliable the test components are as an indicator of the intended test construct.

Step 3: Factor analysis. Exploratory factory analysis was conducted to explore the underlying construct. Factor loadings were examined to explore which component played a greater role in contributing to the total score.

6.4 Questionnaire survey

Questionnaire has been accepted as a widely used and efficient instrument for collecting survey information, and providing structured, often numerical data

(Wilson & Mclean, 1994). It is chosen as the primary instrument in this study for its following advantages: Questionnaire is a cheaper and more cost efficient form of enquiry than interviewing (Weir &Johnson, 1994). It is a time-efficient means of gathering data from a large number of people and is better suited to large-scale surveys (Brown, 2001; Lynch, 1996). Since the present study is intended to collect data from large samples of teachers and students to explore their attitudes on various aspects of the CET-4 and on College English teaching and learning practices, questionnaire will be employed as a major instrument to obtain information from large samples.

6.4.1 Participants

Student participants were from the four sampled universities. Participants from U1, U2, and U3 were all sophomores. They took the CET-4 after they finished one and half a years' English foundation study. Those from U4 were permitted by their university to take the test at the end of Year 1 since students enrolled in this university were supposed to have a better command of English. Teacher questionnaire respondents were from the above-mentioned four universities and other seven universities, ranging from key national universities to ordinary ones.

Varying at academic backgrounds, student respondents were sampled from the following 18 schools or departments of four universities: 1) the School of Journalism and Mass Medium Communication, 2) the School of Law, 3) the School of Economics and Management, 4) the Department of Chinese, 5) the Department of Applied Social Sciences, 6) the Department of Archeology, 7) the Department of History, 8) the School of Life Sciences, 9) the School of Chemical Engineering, 10) the School of Information Science and Engineering, 11）the School of Electronic Engineering, 12) the School of Architecture, 13) the School of Geology Engineering, 14) the School of Software Engineering, 15) the Department of Computer Science, 16) the Department of Mathematics, 17) the Department of Physics. 18) the Department of Chemistry.

6.4.2 Instruments

Two questionnaires were designed for teachers and students respectively. After the preliminary study, both questionnaires were revised and some new variables were incorporated. Given these modifications, ten students and four teachers were invited to complete them before their large-scale administration in order to avoid unanticipated problems and misinterpretation. Based on their feedback, tiny changes were made including further polishing words, and adjusting the layout to make the questionnaire more explicit and accurate.

A standard selected response format was adopted including multiple choice questions, 5-point Likert scales and order-ranking items. The majority of items were designed on a 5-point Likert scale of agreement or frequency. The content and structure of the questionnaires will be detailed in the following part.

Student questionnaire

Quite a number of studies have proved that learner variables such as personality and attitude factors, language aptitude, and learner strategies may interact with test task characteristics in determining test takers' test performance (Bachman, 1990; Hawkey, 1982; Hughes-Wilhelm, 1999; Purpura, 1999; Skehan, 1989; Spolsky, 1989). Therefore, drawing on previous literature, the student questionnaire (SQ) included students' demographical information and background features as the introductive items: gender, age, major, and CET-4 scores. The main body part covered items related to learner variables (motivation, anxiety, learning strategies, extra-curriculum exposure to English, test preparation approaches, test-taking strategies). Some items were adapted from the questionnaires in previous studies (Cheng, 1998; Green, 2007; Hawkey, 2007; Purpura, 1999; Qi, 2004).

There were 35 scales of items in SQ (see Appendix A), which could be classified into three sections as displayed by Table 6.1: students' general evaluations of CET-4 test quality, their test-taking activities (A1-A20), test preparation activities (B1-B12), and evaluations of College English teaching and learning (C1-C7).

Table 6.1 Structure and content of the student questionnaire

Section A	About the CET-4	A1-A20
	Students' evaluations of various aspects of the CET-4 ● Test difficulty ● Demand for students' overall English proficiency ● Test design & content	 A1.1-A1.13 A2.1-A2.13 A8-A14
	Motivations to take the CET-4	A3.1-A3.4
	Decisions made on CET-4 results	A4, A21
	Attitudes on score reporting, the CET-4 SET	A5, A13,A14
	Factors affecting students' listening performances Factors affecting students' reading performances Factors affecting students' overall CET-4 performances	A15.1-A15.5 A16.1-A16.6 A19.1-A19.6
	Test taking strategies for listening comprehension Test taking strategies for reading comprehension	A17.1-A17.5 A18.1-A18.8
	Influences of CET-4 scores on students	A20.1-A20.6
Section B	About the CET-4 preparation	B1-B12
	Effect of test preparation	B1,B2, B5,B6
	Test preparation activities in class	B7.1-B7.4, B11
	Students testing preparation activities ● Time spent on test preparation activities ● Semester to start test preparation ● Use of mock and authentic test papers ● Test preparation methods	 B8.1-B8.10 B9 B3, B4, B10 B12
Section C	About College English teaching and learning	C1-C7
	Motivation for College English learning	C1.1-C1.6
	Extra-curriculum exposure to English	C2.1-C2.7, C5
	Teachers' activities in normal teaching classes	C3.1-C3.8
	Problems existing in College English teaching	C4.1-C4.5
	Evaluations of importance of langue skills Students' self-assessed English proficiency	C6 C7

Teacher questionnaire

The teacher questionnaire (TQ) shared similarity with the SQ in terms of structure and themes (see Appendix B). Teachers' personal particulars were collected covering demographic information such as gender, age, academic qualifications, professional qualifications, years of teaching experience, student type, and weekly teaching loads. The main body part consisted of four sections (see Table 6.2): Section A (A1-A17) explored teachers' perceptions of the CET-4 reform and their evaluations of various aspects of CET-4. Section B (B1-B8) examined teachers' test preparation activities. Section C (C1-C8) investigated teachers' perception of the College English teaching and their classroom teaching activities. Section D

(D1-D6) asked teachers to evaluate students' learning behaviors and factors affecting their CET-4 performances. There were altogether 40 scales of items.

Table 6.2 Structure and content of the teacher questionnaire

Section A	About the CET-4	A1-A17
	Teachers' perceptions of and attitudes to the current CET-4 • Reasons behind the CET-4 reform in 2006 • Perceptions of the changes in the current CET-4 • Possible changes of the CET-4 reform may bring about to teaching	 A1.1-A1.7 A2.1-A2.8 A3.1-A3.6
	Evaluations of the influence of the CET-4 since its inception	A4.1-A4.5
	Teachers' aspects influenced by students' CET-4 performances	A5.1-A5.6
	Teachers' evaluations of various aspects of the CET-4 • Test difficulty • Demand for students' overall English proficiency • Test design & content • Score reporting, CET-4 SET	 A6.1-A6.13 A7.1-A7.13 A8-A14 A15-A17
Section B	About the CET-4 preparation	B1-B7
	Effect of test preparation	B1, B2,
	Test preparation activities	B3.1-B3.4 B5,B6, B8
	Evaluations of doing mock and authentic test papers	B4.1-B4.2, B7
Section C	About College English teaching practices	C1-C10
	Teachers' perceptions of status quo of College English teaching • College English teaching objectives • Problems existing in College English teaching	 C1, C2 C3.1-C3.5
	Teachers' classroom teaching activities • Factors affecting teachers' teaching • Teaching adjustments at the closer time of CET-4 administration • Activities in the normal teaching classes • Typical class size • Satisfaction with textbooks	 C4.1-C4.8 C5.1-C5.4 C6.1-C6.8 C9 C10
	Teaching of test taking strategies for listening comprehension Teaching of test taking strategies for reading comprehension	C.7.1-C.7.5 C.8.1-C.8.5
Section D	Students' learning and testing behaviors from teachers' viewpoint	D1-D6
	Students' motivations for College English learning	D1.1-D1.6
	Students' motivations to take the CET-4	D2.1-D2.4
	Students' testing preparation activities outside of class	D3.1-D3.10
	Factors affecting students' listening performance Factors affecting students' reading performance Factors affecting students' overall CET-4 performance	D4.1-D4.5 D5.1-D5.6 D6.1-D6.6

6.4.3 Backing collection for questionnaires

Student questionnaire

The distribution of SQs consisted of two rounds. In the first round, SQs were distributed to three universities (U1, U2, U3) within the following week of CET-4 administration on 19 December, 2009. In the second round, SQs were distributed to students in U4 two days after they took the CET-4 on 19 June 2010. As mentioned earlier, students in this university were permitted to take the CET-4 at the end of Year 1. The main purpose to choose U4 is that students in U4 at large represent CET-4 test takers with higher English proficiency and samples from this university also serve as the norm group established by the NCETC.

Normally there was still one teaching week after CET-4 administration in these sampled universities. The class periods were for students' self-reviewing and question-consultation. Before the questionnaire circulation, I had contacted the English teachers, who also informed their students of the survey purpose. With the agreement from teachers and students, I went to each classroom and conducted all the collection procedures in person. The class size ranged from about 30 to 50 students. Before the questionnaire distribution, I explained the research purpose and reassured students of confidentiality to encourage their cooperation and true responses. During the process, I occasionally reminded them of the questionnaire structure and different question types in order to avoid any invalid answers incurred by their misunderstanding or ambiguity. It took about 15 to 20 minutes for students to complete the SQ in their class. Moreover, if students were willing to reveal their test scores for research use, they could either write down their student numbers or test taker Identity Numbers so that I would obtain their CET-4 scores from the University Academic Affairs Office.

About 900 student questionnaires were distributed to respondents from the four sampled universities and 817 SQs were returned. The response rate was 90.8%. After data screening, 753 questionnaires were kept as valid, of which 460 questionnaires were matched with respondents' CET-4 scores.

Teacher questionnaire

The teacher questionnaires were circulated in January 2010 via three channels: at teacher meetings, at the CET-4 marking center, and via emails. 1) Some universities held routine meetings at the end of the semester. With the agreement of the program leaders, I explained the purposes of the research in person and teachers cooperatively completed the questionnaires in meetings. 2) When I participated in the CET-4 essay marking, the questionnaires were distributed with the help of supervisors of different marking groups. 3) I sent the TQ via emails to program leaders in other universities, who forwarded the questionnaires to other teachers. After completion, those teachers directly sent them back to my email box.

A total number of 200 TQs were distributed to teachers from 11 universities, and 139 TQs were returned with the response rate being 69.5%. Among the returned TQs, 128 questionnaires were considered valid.

6.4.4 Analysis of backing from questionnaires

***Step 1: Descriptive analysis*.** The valid SQs and TQs were input into SPSS 16.0 for analysis. Statistical results of all the variables including frequencies, means, and standard deviations, kurtosis, skewness were calculated.

Prior to survey analysis, it is necessary to check the internal-consistency reliability of a questionnaire, which refers to the consistency of the answers to questions within a single form of a survey administered on a single occasion. The most commonly reported internal-consistency reliability is Cronbach alpha. It provides an accurate internal-consistency estimate, and it can be used with answers that are coded dichotomously or are on a scale, which makes Cronbach alpha reliability flexible compared to other methods for estimating survey internal-consistency reliability. Conventionally, a value of .70 to .80 is regarded as an acceptable value for Cronbach alpha. However, for a survey designed with subsections that measure distinctly different things, high reliability for the whole survey may not be desirable since the indication is that the subsections might not be as different as the designer initially thought (Brown, 2001). For survey instruments that have distinct subsections, examining the reliability of each of the subsections is much more important than calculating the reliability for the survey instrument as a whole (Brown, 2001).

The overall reliability of the SQ was .840 with reliabilities at section level being .791, .805, and .765, while the overall reliability of the TQ is .953 with reliabilities at sections level .922, .772, .853, and .882. The above results indicate that both questionnaires are reasonably reliable and have a high level of internal consistency.

Tables 6.3 and 6.4 present demographic information of student and teacher respondents. Detailed findings from other variables will be reported as the backing evidence for the corresponding claims and warrants in the following Chapters.

As shown in Table 6.3, among the 753 student respondents, 50.3% of students were female and 49.7% male. Most of students (76.4%) ranged between 19 and 20 in ages. Classified as two broad categories in terms of their majors, 37.7% of students belonged to the Humanity and Arts majors, and 62.3% Science and Engineering majors.

As Table 6.4 shows, 88.3% of the teacher respondents were female and over 90% of them were aged between 26 and 45. In terms of academic qualifications, the largest cohort (68.8%) had Master’s degrees and 30.5% held Bachelor’s degrees. The majority of teachers (71.1%) were lecturers. About half of the teachers had 6-10 years of teaching experience, and 37% of them had more than 10 years of experience. As for student types they taught, half of the teachers were teaching freshmen and the other half were teaching sophomores. More than 70% of them had 9-12 workloads per week.

Table 6.3 Demographic information of student questionnaire respondents

Items	Variables	Percentage (%)
Gender	Male	49.7
	Female	50.3
Age	Below 18	0.4
	18	5.4
	19	37.2
	20	39.2
	21 or above	17.9
Subjects	Humanity & Arts	37.7
	Science & Engineering	62.3

Table 6.4 Demographic information of teacher questionnaire respondents

Items	Variables	Percentage (%)
Gender	Male	11.7
	Female	88.3
Age	Below 25	0
	26-35	65.6
	36-45	29.7
	46-55	3.1
	56 or above	1.6
Academic Qualifications	Below B.A.	0
	B.A.	30.5
	M.A.	68.8
	PhD.	0.8
Professional Ranking	Teaching Assistant	14.1
	Lecturer	71.1
	Associate Professor	14.8
	Professor	0
Years of teaching experience	less than 5 years	18.1
	6-10	45.7
	11-20	29.1
	21-30	5.5
	31 years or above	1.6
Student type	Freshmen	59.4
	Sophomores	40.6
Workload per week (class periods)	less than 8	1.6
	9-10	32.0
	11-12	39.8
	13-14	7.0
	15 or above	19.5

Step 2: Inferential statistical analyses. All the variables were first examined for missing data, outlier cases, normality, multicollinearity and singularity in order to check the distribution and assumptions for more advanced statistical analyses. Even though some sub-items had been grouped together, exploratory factor analysis were conducted with certain items to see whether the results would be in accordance with the hypothetical underlying traits or factors. Since 460 students both provided their CET-4 scores and responded to the questionnaire, their data underwent correlation and multiple regression analyses to examine the relationship between students' perceptions of the test and their test performances. In addition, independent-t tests were performed to compare the means of some overlapping items in both TQ and SQ to explore any attitudinal and behavioral differences in the hope of producing revealing findings.

6.5 Interviews

Data generated from the interviews and the questionnaires are both self–reported data. Each has advantage over the other in certain aspects. Questionnaires enjoy the advantages of being more reliable, anonymous, and economical than interview in terms of time and money. However, their merits are counterbalanced by disadvantages of low response rate and hasty answers (Cohen, Manion & Morrison, 2000). Brown (2001) further indicates that questionnaires are relatively mechanical, artificial, and impersonal in comparison to interviews. They have to be simpler and clearer for there is no opportunity for additional clarification and explanations as would be the case in an interview. In comparison, interviews allow more flexibility than questionnaires. The interviewer can probe more information after a question is answered, and may be able to build a rapport that will help to keep the interviewees interested and motivated in order to obtain rich and spontaneous data (Brown 2001). In addition, interviews can be used to gain insights into questions and topics that have not been predicated in advance and can be pursued and elaborated (Lynch, 1996). Thus, interviews were conducted after the preliminary analysis of questionnaire surveys to help the interpretation of statistical results and obtain in-depth data.

The focused group interview was adopted in this study for its practical and organizational advantages. It is a highly effective data-gathering tool in that a number of people who share specific characteristics of interest to the research are assembled in the focused group, and questions are posed or materials are presented to elicit their reactions. A focused or semi-structured interview usually has an interview guide developed around a list of topics. According to Lynch (1996), the guide acts as a checklist, to make certain that each interview covers the same information, and to allow the interviewer to make efficient use of time and to be systemic and complete across interviews. While the pre-determined topics center on the research questions, the interactive questioning and discussions allow enough flexibility for interviewees to develop areas of concern or volunteer unpredicted content (Minichiello et al., 1995; Weir & Roberts, 1994). In addition, group interviews enable researchers to collect data in a quicker and more efficient way compared with individual interviews. Above all, they provide an effective mode to elicit a wide range of concerns, varied views and allow interviewees to interact with each other. The stimulation of the group members and their reactions to each other' opinions and challenges can result in more revealing responses than a series of individual interviews might make possible (Anastas, 1999; Cohen, Manion & Morrison, 2000; Minichiello et al., 1995).

Therefore, an interview guide was designed for the present study with key prompts related to interview questions and was distributed to interviewees in advance. These guided questions were based on but not limited to the general aspects such as their perceptions of the whole test paper in terms of their content, test methods, the ability that test components intend to measure, their teaching and learning practices, merits and demerits of the CET-4.

6.5.1 Participants

Interview participants were selected from both student and teacher questionnaire respondents. A total number of 30 students with different English proficiency levels and 30 teachers from four sampled universities were recruited. In order to have a more accurate understanding of universities' policies and reasons underpinning their practices related to the CET-4, I attempted to contact administrators from Academic Affairs Offices of the four sampled universities. However, administrators in U3 declined the interviews with the reasons of busy schedule or inconveniences to reveal their universities' regulations, while administrators in other three universities accepted my requests for individual interviews.

6.5.2 Backing collection for interviews

The focused group interviews were conducted from June to July 2010 with student and teacher participants respectively. Before each interview, I briefed the respondents the purpose of my research and got their consents for recording. The interviews were all conducted in Chinese and audio-recorded, lasting from about 20 to 30 minutes. The interviews with administrators were unstructured but were to confirm their policies. They declined being audio-recorded but permitted note taking.

6.5.3 Analysis of backing from interviews

After listening to some interview data and finding no much new information coming up with, I decided not to transcribe and code all the interview data. First, it would be too time-consuming to transcribe about 60 informants' responses in Chinese and then translate them into English. Second, since the major purpose of interviews was to clarify the obscure points, confirm or reject some findings from statistical analysis of questionnaire surveys and students' test scores, questions in the interview guide centered on the research questions and to a large extent overlapped with themes of questionnaire surveys. Given the above reasons, I listened to all the recordings but only interpreted and translated either novel or contradictory information, which would be quoted to generate more in-depth and revealing findings as further backing evidence.

6.6 Test content analysis

Content validity is the representativeness or sampling adequacy of the content of a measuring instrument, which is often determined on the basis of expert judgment (Burns, 2000). This section first reviews the theoretical and operational underpinnings related to content validity, and then proposes a modified and feasible framework for content analysis in this study.

6.6.1 Rationale for test content analysis

Content validity is concerned with whether or not the content of the test (usually including the language skills, structures, etc.) is sufficiently representative and comprehensive for the test to be a valid measure of what it is supposed to measure (Henning, 1987; Hughes, 2003). Messick (1988) thought that content-related inferences are inseparable from construct-related inferences. Bachman (1990) also indicates demonstrating that a test is relevant to and covers a given area of content or ability is a necessary part of validation. Hughes (2003, p.27) stresses the importance of content validity from the backwash perspective: a test in which major areas identified in the specification are under-represented is likely to be inaccurate and to have a harmful backwash effect, since areas that are not tested are likely to become areas ignored in teaching and learning. Therefore, the present study included test content analysis into the research design to help probe into the construct validity of the CET-4.

Messick (1988) stated that content relevance and representativeness of assessment tasks can be addressed by means of job analysis, curriculum analysis and domain theory, which are traditionally appraised by expert professional judgment. Alderson, Clapham, and Wall (1995) echoed Messick’s idea and suggested the content of a test should be analyzed and compared with a content statement which may be the test’s specifications, a formal teaching syllabus or curriculum. Therefore, the first step was to determine what documents were required as the basis for such a comparison. Since test specification is often confidential for internal purposes of the testing committee, test syllabus, as a document for the public, have to be used as an alternative given that it is derived from test specification. In addition, the CET-4, as a criterion-related norm-referenced test, is guided and designed based on the CECR with the purpose to examine and promote College English teaching effect, so it definitely should be treated as an important reference as well. Therefore, the CECR and the CET-4 Syllabus would serve as indispensable and authoritative documents for such a comparison and provide the basis for judgments on content validity.

6.6.2 Framework for content analysis

The necessity for establishing a framework to facilitate experts' evaluation has been stressed by Alderson et al. (1995), who suggest that some data collection instrument should be created and judges should be given a list or some precise indications of the aspects of the test which are to be considered. Therefore, the second step was to specify what facets or test task characteristics should be examined in this framework.

Messick (1988) held that what is judged to be relevant and representative of the domain is not the surface content of test items or tasks but the knowledge, skill, or other pertinent attributes measured by the items or tasks. Bachman (1990) maintains that examining content relevance also requires the specification of the test method facets (TMF). Some empirical studies were conducted in accordance with their ideas. In Alderson's study with Lukmani (1989), judges were provided with a list of skills supposedly being tested by a set of test items, and asked to indicate against each item which skill or kills the item tested. The study by Bachman, Davidson, Ryan and Choi (1995) provided another example in which rating scales were created for a group of experts to rate every task in terms of its task characteristics and the areas of language ability that they believed the task measured (Bachman, 2004). The two studies indicate that regardless of different instruments adopted, experts need to be recruited as judges and provided with a framework to guide their evaluation. The resulting judgments then can be pooled to arrive at an estimate of content validity.

While Bachman's TMF framework, later evolving as task characteristics framework (Bachman & Palmer, 1996, 2010) offers an exhaustive guide for empirical research, the taxonomies in the five categories of TMF framework are too meticulous for the present study to follow rigorously. In addition, categories of testing environment and test rubrics of the CET-4 have been discussed in Chapter 2 (see section 2.4.4). Thus, the nature of the input the test taker receives and the nature of the expected response to that input would be the major sources for the framework designing. Another consideration for major focus on these two facets is that any conclusion drawn from this comparison between the test paper and the official documents should be based on certain criteria. It would be more meaningful and pertinent to the research purpose that the comparison was conducted and analyzed within the range of criteria stipulated in both the CECR and the CET-4 Syllabus. Figure 6.2 displays an overview of this framework modified from Bachman and Palmer's (1996, 2010) task characteristics. The input analysis covered the aspects of topics, genres, and readability. The expected response was analyzed by matching the skills coverage listed in the test syllabus.

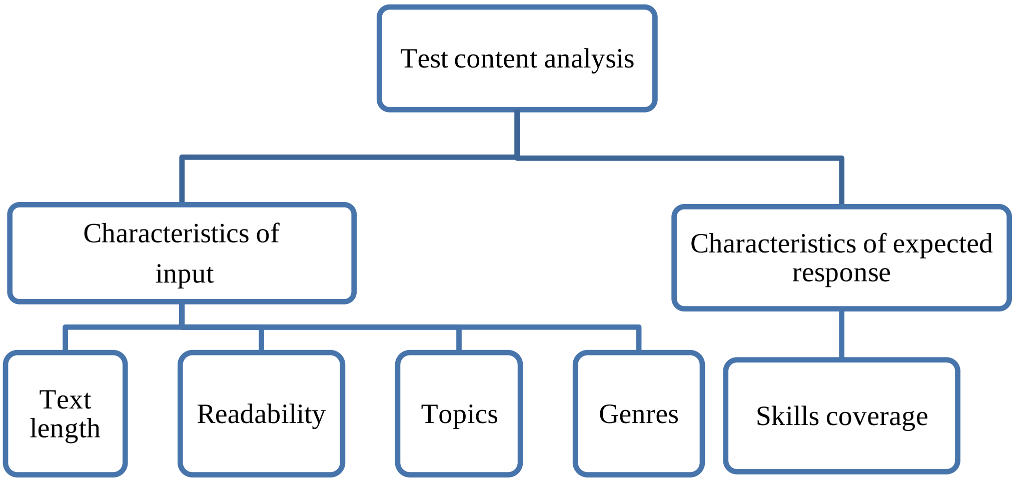

Figure 6.2 An overview of a modified framework for content analysis

The next step was to design tally sheets distributed to experts for analysis, which required reliable coding procedures and valid categories for the classification of the test content data. The validation study on the CET-4 conducted by the NCETC in 1998 provided an authoritative reference and applicable template for the study. The revised 2006 CET-4 Syllabus stipulates that the topics for the CET-4 listening and reading passages are mainly divided into humanities, social science and natural science, and genres are categorized as argumentation, narration and exposition. The expected skills and strategies in the test tasks were examined with reference to skills listed in the CET-4 Syllabus. Drawing on these classifications, the framework was elaborated as more concrete and operational, illustrated by Figure 6.3 below.

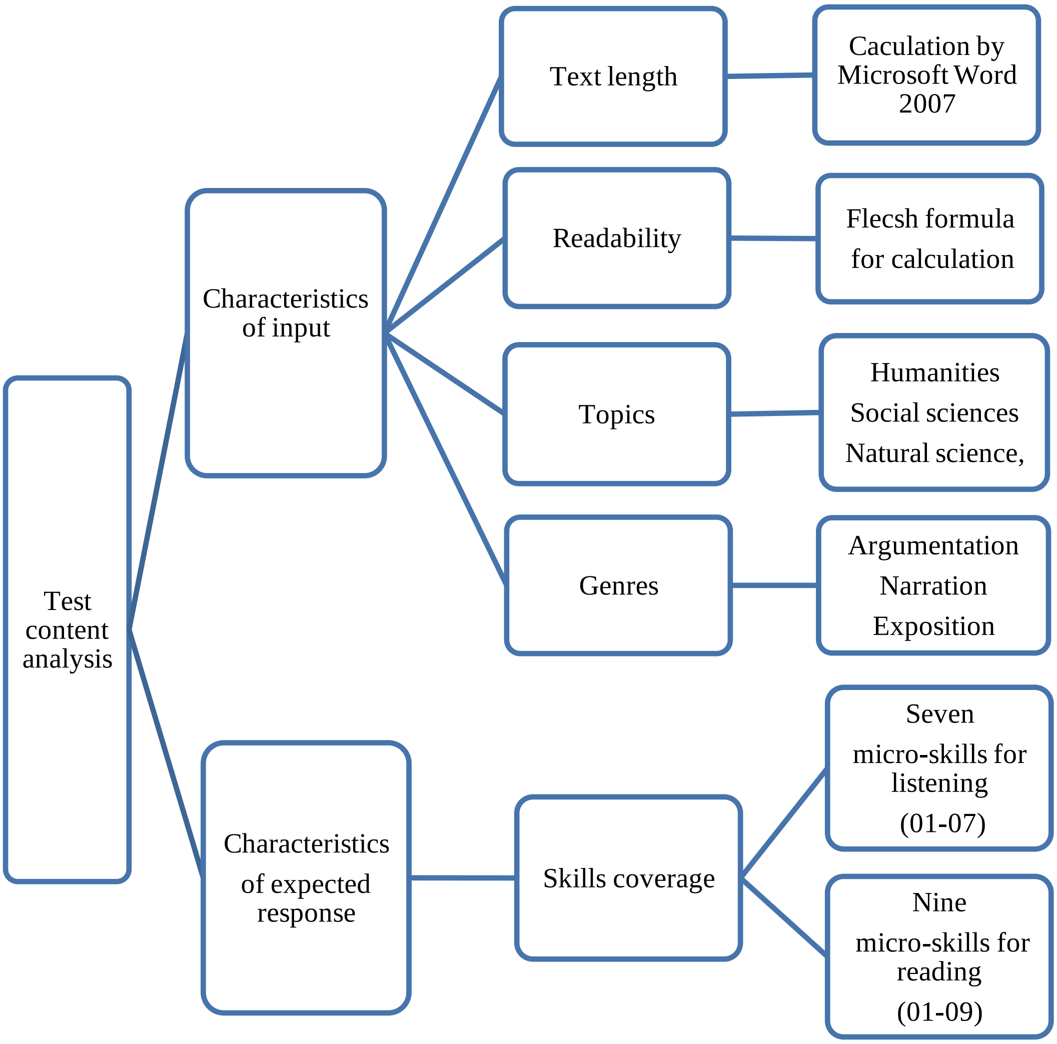
Test content analysis
Characteristics of input
Text length
Caculation by Microsoft Word 2007
Readability
Flecsh formula for calculation
Topics
Humanities
Social sciences
Natural science,
Genres
Argumentation
Narration
Exposition
Characteristics of expected response
Skills coverage
Seven micro-skills for listening (01-07)
Nine micro-skills for reading (01-09)

Figure 6.3 An elaborated framework for test content analysis

6.6.3 Backing collection for content analysis

Since the revised CET-4 was launched nationwide in December 2006, and the test is administered biannually, altogether seven test papers actually used in the past CET-4 administrations (from December 2006 to December 2009) were collected from a commercial publishing house for content analysis. In addition, based on the above framework, the tally sheets were designed. Altogether 11 sets of tally sheets were designed with regard to text length, readability, topics, genres, skills coverage corresponding to both listening and reading components, and one sheet of long conversation turns.

According to Bachman (2004, p.272), collecting the content–related evidence typically involves expert judgment. A group of individuals with expertise in the area of ability being measured, such as researchers, curriculum developers, language teachers and language testers need to read test tasks and make judgments about the ability that each task measures. Thus, two EFL teachers were invited to evaluate the test content with me. They are both associate professors, with Master' degree and more than 10 years of College English teaching experience. Their experience in preparing students for the CET-4 assured their knowledge of the CET-4 formats and contents, as well as their familiarity with the authentic test papers. The three raters were provided with tally sheets, and seven test papers. It took about one month for them to complete all the sheets.

6.6.4 Analysis of test content

For text length, conversation turns, and readability, I conducted all the calculations via Microsoft word 2007 to get the statistics. While pooling our judgments on topics, genres, skills coverage, I followed the following procedures. Each sheet from different raters was coded first. Then I compared each set of sheets and marked all the discrepancies. Since what is needed is to figure out the accurate parameter related to each passage or each item, it was not seriously meaningful to add up the total number of frequencies from each rater and calculate the intra-rater reliability. Therefore, after I marked all the discrepancies, judges assembled, reexamined and discussed passages and items involving discrepancies until consensual judgments were achieved. Finally, the tallies were totaled and the frequencies of these parameters were calculated and summarized in tabular forms.

6.7 Document analysis

Documents, both historical and contemporary, are a rich source of data for education and social research. In conjunction with other data, they can be of importance in triangulation (Denzin, 1989; Punch, 2009). The range of documents may include diaries, letters, essays, personal notes, biographies and autobiographies, institutional memoranda and reports, government pronouncements and proceedings, and policy documents and papers (Jupp, 1996; Punch, 2009).

Documents collected in this study mainly included: publications of the CET-4 administrators and designers, the information from the CET official website, and the 2005 Reform Blueprint for the CET-4 (Wu, 2005) in the press conference, the CECR, and the CET-4 Syllabus. Reviewing and analyzing the above documents was expected to present an overall picture of the test development and design, reveal the test purposes and uses, the major intentions for test reforms, and the underlying theories for test contents and formats.

6.8 Summary

This chapter, based on the existing literature, first justified why a mixed-method approach was appropriate for data collection. It provided a sound rationale for various methods adopted in the research design, and gave a detailed account of the participants, instruments, and procedures in the data collection process.

Both qualitative and quantitative methods were therefore employed to complement each other as well as to enhance the validity and reliability of empirical research (Marton, 1981; Markee, 1994). The employment of a multi-method approach made the AUA more comprehensive and coherent through triangulation of multiple sources of empirical backing evidence. In the following chapters, the major finding generated from quantitative and qualitative approaches will be discussed.

CHAPTER 7

BACKING FOR THE INTERPRETATION CLAIM

7.1 Introduction

Claim 1: The interpretations of students' overall English proficiency are meaningful with respect to the uniform CECR and the CET-4 Syllabus and classroom teaching and learning activities, fair to all the test takers, generalizable to TLU and language teaching domains, relevant to and sufficient for the decisions that are to be made.

This chapter reports on backing evidence collected in support of the interpretation claim which are presented in the above box and to answer the corresponding research question (RQ1): To what extent can the CET-4 serve as an indicator of students' English proficiency? The chapter first presents quantitative analysis of students' CET-4 performances. Results from correlation and exploratory factor analyses are discussed to examine the internal relationship and the test construct. The second part describes content analyses of seven authentic test papers to examine the content relevance and coverage of the CET-4 in accordance with the test syllabus and curriculum objectives.

7.2 Interpretations of test scores

This section seeks answers to RQ1.1. Backing evidence in support of warrants 1 and 2 are to be discussed under the interpretation claim.

RQ1.1: *To what extent does the CET-4 measure the construct to be assessed?*

Warrant 1: The constructs to be assessed by the CET-4 are based on a frame of reference such as the uniform CECR, the CET-4 Syllabus, a needs analysis, or current theory of language use.

Warrant 2: The CET-4 can be interpreted as a useful indicator of the ability to be assessed.

Evidence: In accordance with the teaching objectives and the requirements for students' language skills stipulated in the CECR, the NCETC specifies the test construct and language skills that task types are intended to measure in the CET-4 Syllabus. A few scholarly articles published by members of the NCETC describe the test construct in more detailed and technical terms.

However, no reports from the NCETC can be traced addressing to what extent the construct of the CET-4 is defined with reference to a needs analysis. No detailed reports are available examining the relationships among test components or underlying traits in the CET-4 from statistical perspective. Lack of evidence in this aspect is liable to induce potential rebuttals to this warrant or challenges on construct definition of the test.

Discussions: Thus, the study reported in this section seeks evidence related to construct definition of the CET-4 and investigates to what extent the CET-4 can be interpreted as a useful indicator of the construct. Backing for Warrant 1 can be obtained from official documents like the uniform CECR, the CET-4 Syllabus, and the related publications of the NCETC. Backing for Warrant 2 can be obtained from statistical analyses of students' test performances to explore the internal structure of the test.

In the main study stage, a whole test data package from U1 was collected with permission of the University Academic Affairs Office. The reasons for collecting this test data package were twofold. First, it involved a large number of samples, which could improve the validity of statistical analyses and served as a double check to confirm what had been discovered from the pilot study about test structure and test construct. Second, a proportion of student questionnaire respondents were sampled from these test takers, whose CET-4 performances would be linked to their questionnaire responses for further analyses. Altogether

2982 students took the CET-4 administered in December 2009. After data sorting (see section 6.3.3), a total number of 2692 test data sets were kept as valid. Since the statistical analyses of the 2692 test datasets were the same as those performed in pilot study, and rationales of each statistical procedure and related rules of thumb were explained in Chapter 5, the following part only summarized the major findings.

Table 7.1 presents descriptive analyses of 2692 students' CET-4 performances. The total scores ranged from 347 to 636 with a mean of 461.32 and a standard derivation of 57.69. The values of skewness and kurtosis were within ±1, which were considered excellent in terms of normality. Cronbach's alpha of the whole test is .788.

Table 7. 1 Descriptive statistics of the CET-4 in the main study

CET-4	Min	Max	Mean	SD	Skewness		Kurtosis	
LC	153	239	157.46	28.062	.089	.047	-.395	.094
RC	128	230	167.29	24.010	.125	.047	-.433	.094
CL	44	70	47.80	8.291	-.015	.047	-.472	.094
W&T	85	131	88.76	12.707	-.037	.047	.001	.094
TOT	347	636	461.32	57.688	.159	.047	-.219	.094

Note. LC= listening comprehension; RC= reading comprehension; CL= cloze; W&T= writing & translation; TOT= total score.

As shown in Table 7.2, the correlations among the subtests ranged from .302 to .558, with the correlation coefficient between Listening and Reading being the highest (.558) and that of Cloze and Writing & Translation being the lowest (.302). The correlation between Listening and the total score (.882) and that of Reading and the total (.828) were both across the higher threshold of .80, while the correlations between Cloze and total (.531) and that of Writing & Translation and the total (.681) were moderate.

Table 7. 2 Correlations of the CET-4 in the main study

Subtest	NLC	NRC	NCL	NW&T	NTOT
LC					
RC	.558**				
CL	.368**	.340**			
W&T	.503**	.415**	.302**		
TOT	.882**	.828**	.531**	.681**	

Note. LC= listening comprehension; RC= reading comprehension; CL= cloze; W&T= writing & translation; TOT= total score.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

With the same statistical method applied in the preliminary study (see section 5.2.3.3), Principal components analysis was performed as the extraction method, with oblique as the rotation approach. Table 7.3 displays results of KMO and Bartlett's Test of the CET-4. Table 7.4 indicates that only one factor was extracted , accounting for 56.44% of the variance of the total CET-4 scores. Table 7.5 describes component matrix of the CET-4. Factor loading of the CET-4 ranged between .830 and .633, with the highest loading on Listening Comprehension.

Table 7. 3 KMO and Bartlett's Test of the CET-4 in the main study

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.745
Bartlett's Test of Sphericity	Approx. Chi-Square	239.463
	df	6
	Sig.	.000

Table 7. 4 Total variance explained of the CET-4 in the main study

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.258	56.443	56.443	2.258	56.443	56.443
2	.731	18.286	74.730			
3	.588	14.700	89.430			
4	.423	10.570	100.000			

Table 7. 5 Component matrix of the CET-4 in the main study

	Component
Listening Comprehension	.830
Reading Comprehension	.784
Writing& Translation	.743
Cloze	.633

Note. Extraction Method: Principal Component Analysis

After quantitative analyses of students' CET-4 performances for exploring the test structure and construct, it is essential to address the frame of reference for construct definition so as to better examine the meaningfulness of score-based interpretation. Fulcher (2010, p.96) defines constructs as "the abilities of the learner that we believe underline their test performance, but which we cannot directly observe". Bachman and Palmer (2010, p.43) define construct as "the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task". They further suggest a frame of reference for construct definition, which may include "a course syllabus, a needs analysis of TLU tasks, a theory of language ability, or combination of these" (Bachman & Palmer, 2010, p. 213). The following part will discuss the intended construct of the CET-4 based on this frame of reference.

Since the CECR and the CET-4 Syllabus are the official documents pertinent to the CET-4, it is necessary to see how these two uniform syllabuses address university students' English ability and the intended construct of CET-4. According to the CECR (Ministry of Education, 2004, p.5), the objective of College English is to develop students' ability to use English in an all-round way, especially in listening and speaking. It also delineates the specific requirements at basic, intermediate and higher levels for students' listening, speaking, reading, writing and translation abilities. As noted earlier, the administration of the CET-4 is to serve the implementation of the teaching syllabus. Thus, the objective of College English Test publicized by the NCETC on its official website states that "CET-4 and CET-6 are aimed at measuring precisely college students' comprehensive employment of English and thus play an active role in realizing the objective of college English teaching" (CET, 2011). The above specifications indicate that the CET-4 is intended to measure students' overall English proficiency represented by their performances on skills of listening, speaking, reading, writing and translation. In accordance with the two syllabuses, the CET-4 is designed with four components, namely, Listening, Reading, Integrative part (Cloze), and Writing and Translation. As to students' speaking ability, it is separately measured in the CET-4 SET.

In respect to the above test score analyses, correlations and EFA were conducted on the Score Report Form issued by the NCETC, which consists of a composite score and profile scores from four components (Listening, Reading, Integrative skill, and Writing & Translation). On the one hand, the correlations among the components fell into the range of .30 to .60, implying that skills indicated by the four components were distinct. On the other, the correlations between components and the total ranged from .531 to .882, implying they were not completely independent of each other but related because they all contributed to a general factor underlying the CET-4. Similarly, the exploratory factor analysis extracted only one factor. Based on document review of the two syllabuses, this factor can be interpreted as a general factor of English proficiency encompassing four latent traits. Statistically speaking, results of test score analysis from the CET-4 administered in December 2009 in the main study are similar to those from the CET-4 administered in December 2008 in pilot study, and similar to what has been released by the NCETC about the trialed version of the revised CET-4. It has evidenced again that the post-2006 CET-4 demonstrates better internal structure and measures test takers' overall English proficiency.

Since the most significant change in the current CET-4 is the increasing weight of Listening and the revised listening task types, it is necessary to examine statistics relevant to the Listening part. As noted earlier, the Listening part not only had highest correlation (.882) with the total score but also had largest factor loading, which served as evidence that the current CET-4 lays more emphasis on measuring students' listening ability. Second, the needs analysis conducted in phase I explored university students' needs for English communication in real-life situations, and a four-factor solution was produced by EFA (see section 5.3). Factor 1 accounting for the largest variance was concerned with both formal and informal interactions requiring effective communication skills in English especially speaking and listening abilities. This finding offers another source of evidence that the current CET-4 responds to students' needs to improve listening ability and to the social demands for graduates with higher communicative English proficiency in its 2006 reform

With respect to needs analysis, the NCETC claimed that they conducted a large-scale needs analysis before the launch of the CET-4 in the late 1980s, and accordingly designed the test to meet students' and social requirements for strong reading ability during that period. In terms of the current CET-4 launched since 2006, Jin (2006) declares that the next step in the ongoing reform is to conduct a new round of needs analysis. However, no official report heretofore has been released by the NCETC about students' needs for English proficiency in TLU domain. This may become a potential rebuttal to the underlying reason for reforming the test reform.

Since a theory of language ability tends to serve as the theoretical basis of test construct, it is necessary to investigate how the NCETC defines the construct based on language ability. When it was launched in 1987, the CET-4, under the influence of the psychometric structuralist approach and the componential view of language construct, was designed as a component test composed of several sections assessing the four major language skills, namely listening, speaking, reading and writing (Jin & Yang, 2006). Great importance was attached to measuring test takers' linguistic knowledge and selected response items occupied dominant proportion. With Bachman's proposal of communicative language ability in the early 1990s, the NCETC acknowledged the importance of this language ability model but asserted that linguistic competence should serve as its basis (Yang, 2000; Yang & Jin, 2000, 2001; Yang & Weir, 1998, p.60). Thus, the CET-4 was designed with a combination of both analytical and integrative approach, with certain proportions of discrete point items on separate skills, items on integrative skills and items on communicative language ability. Up to now communicative language ability and communicative teaching approach have been widely accepted. Large-scale and well-established tests like TOEFL iBT and IELTS are designed as communicative tests in response to the worldwide demands for test takers' higher communicative English proficiency. Hence, the 2006 CET-4 reform has kept pace with this trend in the field of language testing by including more communicatively-oriented tasks and constructive response items. However, it is interesting to notice that official documents and most articles published by the NCETC (e.g., Jin, 2004, 2005, 2006; Jin &Yang, 2004; Zhang, 2008) address construct underlying the CET-4 as *yingyu zonghe yingyong nengli* in Chinese (meaning comprehensive employment of English) rather than communicative language ability. The CECR specifies the intended language proficiency as the ability to use English in an all-round way, especially in listening and speaking. The NCETC interprets it as the comprehensive employment of English in their official CET website. The possible reason may be that the NCETC intends to define the construct in non-technical language in order to facilitate test takers' and test users' accurate understanding. Only two articles were traced in which Communicative Language Ability is explicitly used as a technical term to define the construct to be measured by the CET-4 (Jin &Yang, 2006; Yang, 2004). With regard to the current CET-4, the NCETC claims that the reform of the CET is conducted as "a response to the pressing social need for college and university graduates with a stronger communicative competence in English" (Jin &Yang, 2006, p.21) and "the task of the CET designers is to provide a comprehensive assessment of the testees' communicative language ability" (ibid, p.35).

Section summary

When evaluating the construct of the test, we cannot simply accept construct labels test developers used. That is why in previous part correlations and exploratory factor analysis were conducted to explore the internal construct of the test, and documents and scholarly articles were further referred to in our score-based interpretations. Based on the aforementioned construct definitions and preliminary statistical analyses on the structure of Score Report Form, a hypothesized second-order factor model could be posited. This model would demonstrate four latent traits underlying the test performance and a second-order factor of overall English proficiency. It hypothesizes two kinds of relationships. One is that the four latent traits are distinct and a test taker's abilities on the four skills are indicated by their performances on the subsection tasks. The other is that the four latent traits are interrelated and all contribute to one general factor indicating test takers' overall English proficiency. The next step is to verify this model via confirmatory factor analysis. However, without access to subsection scores of different task types under each component, advanced statistical procedures like structural equation modeling (SEM) cannot be employed to reveal the relationship between test components and the general English proficiency that the test is intended to measure, as well as the relationship among the four underlying traits. This is a major limitation of the present study.

By the same token, it would be one source of potential rebuttal on the construct of the CET-4 since the NCETC has not released in-depth empirical evidence in any academic articles or research projects. It is understandable that the NCETC intends to keep test data information confidential, but test developers cannot expect test takers and test users to buy whatever construct label they verbally described without releasing convincing and authoritative statistical evidence. Both ETS and the Cambridge ESOL have taken a lead in this aspect. They sponsor researchers outside test agencies with financial assistance and access to some real test data for studies on TOEFL and IELTS. They post a series of research projects on their websites as supporting evidence for test qualities. A large body of research in turn demonstrates solid base of scientific research and development underlying the tests, which makes the test more credible, objective and widely accepted by test users.

7.3 Content analyses of listening and reading components

This section seeks backing for RQ1.2 and the third warrant under the interpretation claim:

RQ 1.2: *To what extent is the CET-4 representative of the content relevance and coverage in accordance with the test syllabus and curriculum objectives?*

Warrant 3: The CET-4 is meaningful and generalizable for its content representativeness and content relevance in accordance with the test syllabus and curriculum objectives.

Evidence: The CECR set detailed language skills requirements at different levels for students to meet and accordingly the CET-4 Syllabus explicitly describes characteristics of each task type that test takers are expected to perform.

However, no detailed reports are available from the NCETC checking the degree of correspondence between the content of the authentic test papers since 2006 and the task characteristics specified by the CET-4 Syllabus and the requirements set by the CECR. This lack of evidence would become the potential rebuttal to the interpretation claim and would challenge the content validity of the test.

Discussion: Thus, this section examines content validity of listening and reading components in the CET-4 by reviewing the test papers actually used in the past CET-4 official administration to examine whether the test items adequately represent samples of the behaviors or content domains that both teaching and test syllabuses intend to measure. Based on the modified framework of task characteristics proposed by Bachman and Palmer (1996, 2010), which has been discussed in Chapter 6 (see section 6.6.2), the major findings will be reported from two perspectives, characteristics of input covering text length, readability, topics and genres, and characteristics of expected response mainly about listening and reading skills coverage. All the test papers were collected and published by a commercial publishing house in China. Altogether 126 text passages and 490 question items from listening and reading materials in seven test papers, spanning from December 2006 to December 2009, underwent thorough analysis (see Table 7.6). The approach to and focus of analysis on each subcomponent may vary with format and content of different test tasks. The following parts will give a detailed account of the above parameters.

Table 7.6 The number of texts and items of listening and reading components

Components	Task types	Text materials	Question Items
Listening	Short conversations	56	56
	Long conversations	14	49
	Passages	21	105
	Compound dictation	7	70
	Total	98	280
Reading	Fast reading	7	70
	Banked cloze	7	70
	Passages	14	70
	Total	28	210
	Total	126	490

7.3.1 Characteristics of input

Document analyses serve as an important component in the present study, so the CECR and the 2006 CET-4 Syllabus are included as important instruments for content analysis. Since their development and revisions have been thoroughly reviewed in Chapter 2, Tables 7.7 and 7.8 below highlight the related input criteria stipulated in the two syllabuses, serving as a baseline to examine the degree of correspondence in comparison with results of content analysis.

Table 7.7 Listening requirements of teaching and testing syllabuses

Requirements	NCETS (2004)	CET-4 syllabus (2006)
Genre& Topic	Students should be able to follow classroom instructions, everyday conversations, and lectures on general topics conducted in English	Genres: narration, exposition and argumentation, etc Topics: humanities, social science, natural science, etc
Speech rate	Students should be able to understand Special English programs spoken at the speed of about 130wpm	Standardized British or American English presented at a speed of 130 wpm
Length	No specific requirements	Listening passages: 200-250 words Compound dictation: 200-250 words
Readability	No specific requirements	Less difficult level

Table 7.8 Reading requirements of teaching and testing syllabuses

Requirements	NCETS(2004)	CET-4 syllabus (2006)
Genre& Topic	Students should be able to read English newspapers and magazines published in China and understand texts of practical styles commonly used at work and in life.	Genres: narration, exposition and argumentation, etc Topics: humanities, social sciences, natural science, etc
Reading speed	Students should be able to read English texts on general topics at a speed of 70 wpm, and longer yet less difficult texts at 100 wpm.	Fast reading: at a speed of 100 wpm
Length	No specific requirements	Banked Cloze: 200-250 words Reading Passages: 300-250 words Fast Reading: 1000 words
Readability	No specific requirements	Intermediate difficulty for reading in depth, Less difficult level for fast reading

7.3.1.1 Text length

The length of the input is the amount of language that the test taker needs to process. It is very important in reading and will also affect processing difficulty (Alderson, 2000). Text length in the present study refers to quantity of words and was calculated by Microsoft Word 2007. The calculation was conducted on short listening and compound dictation passages, fast reading, banked cloze and careful reading passages.

Listening

According to the CET-4 syllabus, the length of both short listening passages and compound dictation should be between 200-250 words per passage. Table 7.9 lists the word quantity of each passage and the total average. The average length of short listening passages is 239 words and the average for compound dictation is 206, which are both within the range of test syllabus requirement. Further examinations reveal that the total words of short listening passages range between 217(2008/12 A) and 284 (2007/6 C), while the length of compound dictation ranges from 176(2008/12) to 237(2009/12). Only two passages are out of the specified range of text length with the minimum length of 176 words and the maximum of 284 words, while the rest of passages are within the required range of 200-250 words. It indicates that test developers on the whole select passages with appropriate text length and the word quantity of each passage in the listening component did not fluctuate too much.

Table 7.10 presents the number of turns of Long conversations, ranging from five to twelve turns with seven as an average. The test syllabus lists 5-8 turns as the reference criterion. Therefore, this task overall is in line with the syllabus requirement.

However, some inconsistence can be observed as well. Five out of the total fourteen conversations are beyond 8 turns, and Conversation A in test papers of Jun. 2007 and December 2009 even amounts to 12 turns, while two conversations in test paper of June 2008 only have 5 turns. This indicates a subtle instability in the conversation turns, which deserves test designers' attention, since test takers' performance may be affected if the length of conversations exceeds what test takers expect or accept.

Table 7.9 Text lengths of listening passages and compound dictation

Test paper	Listening passages				Compound dictation
Requirements	*200-250*				*200-250*
	A	B	C	Average	
2006/12	229	234	242	235	201
2007/6	241	250	**284**	258	203
2007/12	243	238	228	236	216
2008/6	238	244	253	245	199
2008/12	**217**	237	252	235	**176**
2009/6	225	222	229	225	212
2009/12	237	243	244	241	**237**
Total average	233	238	247	**239**	**206**

Table 7.10 The number of turns in long conversations

Long conversation	06/12	07/6	07/12	08/6	08/12	09/6	09/12	Average
Conversation A	5	12	9	5	7	6	12	8
Conversation B	6	7	6	5	9	7	5	6
Average	6	10	8	5	8	7	9	7

Reading

According to the CET-4 syllabus, the length of fast reading should be about 1000 words, the length of Banked Cloze between 200 and 250 words, and the length of passages between 300-350 words. Table 7.11 shows the word counts of reading comprehension.

The words of fast reading range from 1017 to 1100 with the average of 1023. It did not show abrupt increase or decrease in its word quantity. The average length of

Banked cloze is 240 words, and only one passage (2009/6) is beyond 270 words. The average length of careful reading passages is 352 words. Fourteen passages range from 328 to 362 words. Compared with the word range of 300-350 in the test syllabus, 12 passages fall in the interval of 350-370 words. It can be seen that careful reading passages pose a high demand for students' processing of longer passage input.

Table 7.11 Text lengths of reading comprehension

Test paper	Fast reading	Banked cloze	Careful reading 1st	 2nd	 Average
Requirements	*1000*	*200-250*		*300-350*	
2006/12	**1017**	**220**	351	**362**	357
2007/6	1020	243	350	**362**	356
2007/12	995	224	358	353	356
2008/6	962	235	351	352	352
2008/12	996	228	**362**	353	358
2009/6	**1070**	**272**	351	354	353
2009/12	1100	255	**328**	338	333
Total average	1023	240	350	353	**352**

To sum up, in terms of the text length, the average lengths of both listening and reading comprehension tasks are quite within the length range put forward by the CET-4 syllabus. Text length calculations indicate nice correspondence with the test syllabus. Moreover, the discrepancy between the maximum and the minimum of words is not big, usually around 50 words, which reveals a better consistence in designing and selecting appropriate length of listening and reading materials.

7.3.1.2 Readability

One of the frequently used and widely acknowledged readability formulae is the Flesch formula, which produces a reading-ease score: RE4= 206.835-(0.846 x

NSYLL) – (1.015 x W/S), where NSYLL is the average number of syllables per 100 words and W/S is the average number of words per sentence (Davis, 1984, p.88; Alderson, 2000, p.71). The present study examined the readability via Microsoft Word 2007 that stipulates a similar formula: 206.835-(1.015 x ASL) - (84.6 x ASW). ASL refers to average sentence length and ASW means the average number of syllables per word (http://office.microsoft.com/en-us/word-help/test-your-document-s-readability-HP010148506.aspx?CTT=1#BM1). In order to have a better understanding of these figures, I will first evaluate them referring to the Flesch readability reference table proposed by Yang and Weir (1998), and Gu and Guan (2003) in their studies. Table 7.7 presents the scales of reading-ease score and the yardstick of their corresponding difficult level. It can be seen that the lower the reading-ease score is, the more difficult the text is.

Table 7.12 Scale of reading-ease score (Yang & Weir, 1998; Gu & Guan, 2003)

Scale	0-30	30-50	50-60	60-70	70-80	80-90	90-100
Difficult level	Very difficult	Difficult	Fairly difficult	Standard	Fairly easy	Easy	Very easy

Table 7.13shows the readability of listening components in seven test papers, calculated via the formula stipulated by Microsoft Word 2007. The average readability scores for short listening passages (61.0) and compound dictation (59.6) are at approximately the same level, being standard. The second passage of test paper 2007/12 is the most difficult (46.0), while the second of 2009/6 is the easiest (81.6). In addition, if we examine the average score of short listening passages, we can find that passages of 2007/12 rank as the most difficult (51.0) in the seven test papers, while those in the next administration (2008/6) are the easiest (69.6). Such sharp discrepancies in text difficulty can also be observed in compound dictation of 2008/12 and 2009/6 (48.0, 72.4). Test designers may adjust the difficult level based on test takers' reaction and their test performances, but consistency and degree of adjustment should be taken into consideration so that test difficulty will not fluctuate too much in two successive tests.

Table 7.13 Readability of listening passages and compound dictation

Test paper	Listening passages				Compound dictation
	1st	2nd	3rd	Average	
2006/12	48.0	63.7	76	62.6	64.7
2007/6	61	61.8	71.5	64.8	69.2
2007/12	57.9	**46.0**	49.0	**51.0**	53.5
2008/6	62.4	68.6	77.7	**69.6**	58.7
2008/12	62.7	52.2	46.8	53.9	**48.0**
2009/6	55.8	**81.6**	67.9	68.4	**72.4**
2009/12	64.1	59.7	47.7	57.1	50.9
Average	58.8	61.9	62.4	**61.0**	**59.6**

Table 7.14 summarizes statistics from reading component. The average readability scores for banked cloze (56.1), fast reading (47.8), and careful reading (53.8) vary with their different test methods. It is noteworthy that the readability score of fast reading is in the scale of difficulty and the passage of 2007/12 is very difficult with readability of 23.5, while banked cloze and careful reading are fairly difficult. Both CECR and 2006 CET-4 syllabus stipulate that fast reading should be longer yet less difficult than reading in depth. In this sense, it does not quite comply with syllabus requirements. Since fast reading is designed to check students' skills of locating information and grasping main idea of longer input within limited time, it is not advisable to select materials that are more difficult. Moreover, when fast reading was firstly included in test paper of 2006/12, its readability is 70.0, but that of 2007/12 passage decreased to 23.5. The difficulty level fluctuated from standard to very difficult scale. The gap between the maximum and the minimum is too great to maintain its stability.

Table 7.14 Readability of reading component

Test paper	Banked	Fast	Careful reading		
	cloze	reading	A	B	Average
2006/12	53.4	70.0	49.3	58.0	53.7
2007/6	55.0	45.7	71.5	64.6	68.1
2007/12	42.2	23.5	67.1	46.9	57.0
2008/6	73.7	37.4	43.7	57.6	50.7
2008/12	51.3	66.8	65.4	41.7	53.6
2009/6	64.8	48.2	50.2	54.5	52.4
2009/12	52.2	42.9	53.4	28.7	41.1
Total average	56.1	47.8	57.2	50.2	53.8

Table 7.15 presents classifications of passages based on their readability in order to have an overall picture of the difficult levels of the two components. It can be seen that the readability of listening components are mainly in the scale of 60-70 and 50-60, as standard (32.1%) and fairly difficult (28.6%) level. In order to discriminate students with different listening abilities, 21.4% of passages are at the difficult scale and only 14.3% are fairly easy. While for the reading components, the majority of passages are in the scale of 30-60, being difficult (32.1%) and fairly difficult (32.1%). It is also in accordance with the two syllabuses that listening tasks should be less difficult than reading tasks.

Table 7.15 Readability classifications of listening and reading components

Scale	0-30	30-50	50-60	60-70	70-80	80-90	90-100
Difficult level	Very difficult	Difficult	Fairly difficult	Standard	Fairly easy	Easy	Very easy
Listening	0	6(21.4%)	8(28.6%)	9(32.1%)	4(14.3%)	1(3.6%)	0
Reading	2(7.1%)	9(32.1%)	9(32.1%)	5(17.9%)	3(10.7%)	0	0

After evaluating these tasks with reference to the Scale of reading-ease score, I further take a diachronic approach to make a comparison with statistics of passages before 2006 CET-4 reform revealed in previous studies. The 1998 CET validation study investigated the content validity of short listening and reading passages from 1987 to 1996, revealing their average readability scores of 75.0 and 57.7 respectively (Yang &Weir,1998). Li (2009) and Jiang (2009) further investigated readability of listening and reading passages of test papers from 1997 to 2006. Table 7.16 presented scores of the present study with those in the above two studies. The three studies are also in accordance with three stages of CET-4 development, as mentioned in section 2.2 of Chapter 2. The table reveals that listening comprehension has been increasingly difficult until now. In terms of reading comprehension, the difficulty level increased from the first stage (57.7) to the second (51.3) but decreased a little bit (53.8) in the third stage. The changing trend in readabilities of listening and reading components in the past years are in line with the changing focus of the CET-4 shifting from its traditional emphasis on examining students' reading ability to examining students' listening ability. Accordingly, the revised CET-4 modified its test content and format by increasing weight of listening component and diversifying listening task types.

Table 7.16 Readability of listening and reading passages from 1987 to 2009

	Stage 1(1987-1996) (Yang &Weir,1998)	Stage 2(1997-2006) (Li,2009; Jiang ,2009)	Stage 3(2006-2009) **(**The present study**)**
Listening passages	75.0	63.2	61.0
Reading passages	57.7	51.3	53.8

7.3.1.3 Genre

Alderson (2000) believes that certain topics are associated with certain types of texts. What causes difficulty in text is less the actual content than the way the text is written: its style, or the features that make one text different from another, and that gives rise to a number of different classifications of text type. Based on various levels of generality or specificity, text genres can be described in many different ways (Johns, 2002). Genres can be discussed from discourse and nondiscourse perspectives, or described in terms of daily-used texts or instructional genres. Macro-genres are defined as encompassing text types identified as expository, narrative, persuasive (Grabe, 2002) or as narrative, recount, argument, and report (Hyland, 2002). In view of the complexities of genre classifications by different genre schools, the present study based its genre classifications on both the CECR and the CET-4 syllabus, which clearly stipulate that genres of selected materials mainly include three categories: narration, exposition and argumentation.

It is necessary to first touch on basic definitions and major features of the three genres to make classification criteria more explicit and accurate. Narration describes personal experience in time sequence. It is often synthesized with argumentation since feelings, ideas, and inspiration are often revealed or concluded by the author. If the passage largely narrates personal experience, it will be grouped into narration. Exposition intends to notify or explain something to readers. In term of its general structure, a question or a topic is usually proposed from an example for discussion, then its explicit and implicit causes are analyzed and finally it comes up with a conclusion or a solution. Its language use contains more objective terms and less subjective judgment vocabulary. Argumentation aims to persuade readers to agree with the opinions proposed in the passage. Its structure features usually contain a controversial viewpoint, demonstration and a conclusion. Deduction and induction are the two major reasoning approaches. Based on the above classifications, the present study examined genres of short listening passages, compound dictation, and three task types of reading component.

Listening

Table 7.17 presents genres of short listening passages and compound dictation. Ten out of the total twenty-one short listening passages belong to narration, while none of the seven compound dictation passages is this type of genre. Five out of the seven compound dictation passages are of exposition. If we examine the total percentage of both task types, we can see that exposition (42.9%) occupies the largest proportion, followed by narration (35.7%) and argumentation (21.4%).

Table 7.17 Genres of short listening passages and compound dictation

Test paper (2006/12-2009/12)	Narration	Exposition	Argumentation
Short listening passages	10	7	4
Compound dictation		5	2
Total	10 (35.7%)	12 (42.9%)	6 (21.4%)

Similar diachronic approach was adopted to make a comparison with statistics of short listening passages revealed in previous studies on the old CET-4 (see Table 7.18). Through the three stages of CET-4 development, narration can be identified as the most adopted genre except in stage 2 in which exposition topped with 54.9% of distribution, while argumentation occupies the least proportion except ranking as the second in stage 1.

Table 7.18 Genres of short listening passages from 1987 to 2009

	Stage 1(1987-1995) (Yang &Weir,1998)	Stage 2(1996-2006) (Li,2009)	Stage 3(2006-2009) **(**The present study**)**
Narration	19(42%)	18(35.3%)	10(47.6%)
Exposition	12(26.7%)	28(54.9%)	7(33.3%)
Argumentation	14(31.1%)	5(9.8%)	4(19.0%)
Total	45	51	21

To sum up, genres of listening passages are appropriately distributed. First, each genre is representative in certain proportion of passages, which is in line with the syllabus requirements. Second, the genre distribution is overall reasonable and scientific. According to Alderson (2000), expository texts are generally considered harder to process than narrative texts, whose conventionalized macro-structures associated with stories seem to facilitate comprehension by allowing readers to quickly construct a model of the text. This may justify why narration is the mostly used genre in short listening passages. In addition, the CECR states that listening materials should be less difficult than reading, so test takers' abilities to grasp main idea and specific details are more densely examined than their inferring and deducting abilities in this part. Narrative passages can better facilitate designing items targeting at checking students' these abilities. In other words, it is reasonable to control the number of argumentative passages in listening comprehension, because this genre poses higher requirements for students to distinguish conflicting viewpoints and infer speaker's attitudes, and its structure characteristics make the passages more difficult. This may help understand why argumentation is the least adopted genre. With reference to compound dictation, which requires students' productive ability, its difficulty level is supposed to be improved a bit. In particular, the three sentence completion items ask students to summarize what they hear in their own sentences. In order to prevent students from easily inferring the plot of narrative event, more expository materials are selected to achieve the end.

Reading

With regard to reading component, an interesting pattern can be identified from Table 7.19. All the passages fall into categories of exposition (60.7%) and argumentation (35.7%) except one banked cloze passage belonging to genre of narration (3.6%).

Table 7.19 Genres of banked cloze, fast reading and careful reading passages

Test paper (2006/12-2009/12)	Narration	Exposition	Argumentation
Fast reading		5	2
Banked cloze		6	
Careful reading	1	6	8
Total	1 (3.6%)	17(60.7%)	10 (35.7%)

Further comparison was made between the present study and previous studies. Since both banked cloze and fast reading are newly added elements, Table 7.20 only presents statistics of careful reading passages. Before 2006, each test paper contained four passages, while the number has been reduced to two since 2006. Contrary to genre distribution of short listening passages, narration is the least used genre. In stage 1, 59 out of 60 passages were argumentation genre according to the figure released by the NCETC in their 1998 validation studies (Yang &Weir, 1998). It seemed that the NCETC was aware of this problem and then included more expository materials during the second decade of CET-4 development, in which exposition was the dominant genre, followed by argumentation (Jiang, 2009). In the new CET-4, exposition still takes up 42.9% of the total genres while the rest of 57.1% is distributed in argumentation. Although the present study only examined seven test papers within about four years, it is likely that exposition and argumentation will share similar proportion of total genre in the potential trend.

Table 7.20 Genres of careful reading passages from 1987 to 2009

	Stage 1(1987-1995) (Yang &Weir,1998)	Stage 2(1996-2006) (Jiang ,2009)	Stage 3(2006-2009) (The present study)
Narration	1 (1.7%)	5(5.4%)	0
Exposition	0	56(60.9%)	6(42.9%)
Argumentation	59(98.3%)	30(32.6%)	8(57.1%)
Practical writing		1(1.1%)	
Total	60	92	14

In summary, when we examine the above two tables, it is easy to make a tentative conclusion that the genre distribution of reading component is unbalanced and irrational. A surface impression is that narration seems to be largely ignored while exposition and argumentation are favored in the CET-4 material selection. However, some reasons may account for this genre distribution. According to the CECR, fast reading, requiring students to grasp main idea and locate the specific details, should be longer yet less difficult. Exposition genre provides test designers with diverse materials and a wide range of topics while holding less difficult level than argumentation. Compared with narration that is too familiar for test takers predict, frequent adoption of exposition in fast reading makes the test materials diverse and appropriate at the difficult level. In addition, in real life situations, when students tend to employ skimming and scanning skills for either casual reading or professional reading, the materials they handle are highly likely to be a genre of exposition. In this sense, the genre distribution also represents reading samples in target language use domain. Moreover, it also facilitates the design of banked cloze, for abundant content and topics help to diversify the vocabulary bank. When test takers are required to choose from the word bank for gap filling, their vocabulary knowledge can be examined in broad and profound dimensions. However, in terms of careful reading, the number of argumentative passages outnumbered that of exposition. This is because more in-depth reading skills such as inference and deduction can be better examined in the argumentation genre. Careful reading with its increasing difficulty can also help discriminate students with different reading abilities.

7.3.1.4 Topic

Alderson (2000) suggests that good tests of reading and good assessment procedures in general will ensure that readers have been assessed for their ability to understand texts in a range of different topics. Therefore, an examination of the CET-4 topic areas is necessary and meaningful in content validation. Traditionally, the NCETC in their validation report classified topics of listening comprehension into four categories: daily life, society and culture, natural science, and biomedicine, and topics of reading comprehension into three categories: humanity and management, science and technology, and biomedicine (Yang & Weir, 1998). In 2006 CET-4 syllabus, the topic range is broadly generalized within humanities, social science, and natural science, etc. In order to reduce inaccuracy induced by ambiguous or different understandings of topic boundaries, a meticulous taxonomy of topics was further made to facilitate our grouping with explicit and uniform criteria (See Figure 7.1).

In the present study, passages related to linguistics, literature, arts, religion, history, philosophy belong to humanities. Social science as an umbrella term covers a wider range of fields, such as economics, psychology, political science, sociology, criminology, anthropology. Natural science mainly includes biology, medicine, earth sciences, physics, chemistry, astronomy, etc. In terms of short and long conversations, it is difficult to classify them given their short length especially the one-turn conversations, but a general description of their topic domains will be presented first.

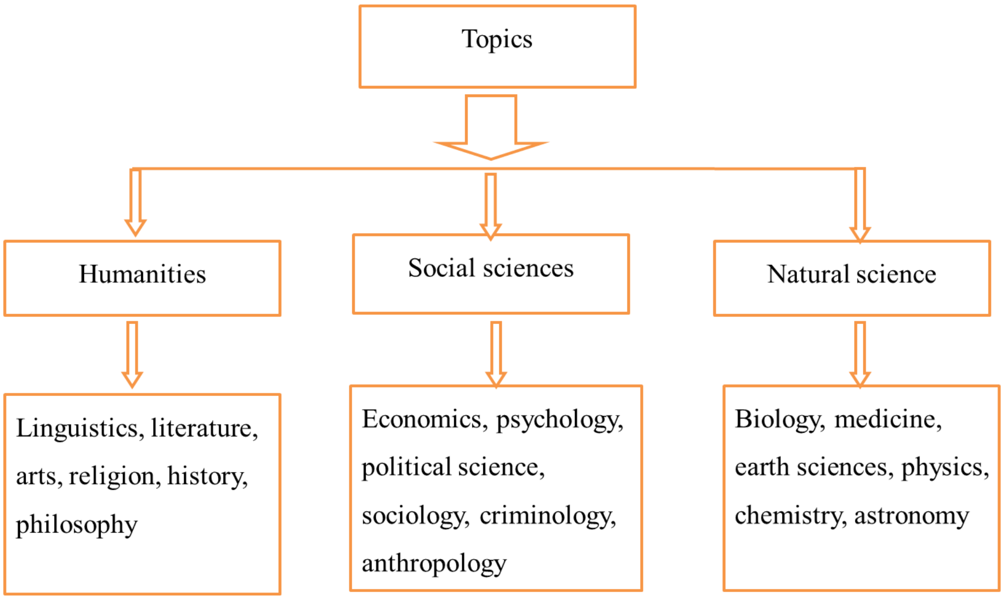
Topics
Humanities
Social sciences
Natural science
Linguistics, literature, arts, religion, history, philosophy
Economics, psychology, political science, sociology, criminology, anthropology
Biology, medicine, earth sciences, physics, chemistry, astronomy

Figure 7.1 The taxonomy of topics of listening and reading passages

The majority of short conversations are related to daily life, covering a variety of settings such as shopping, vacations, hobbies, going to dining, buying movie tickets, watching TV programs, talking about weather and persons, etc. Topics related to academic study and school life are concerned with attending seminars, discussing course selection or contents with students and professors, doing assignments, taking examinations, and borrowing books from library. These topics are familiar to students, since they are frequently engaged in these small talks in real life situations and are exposed to similar academic contexts.

The long conversations deal with travel planning (2006/12 LC1), hotel reservation (2007/6/LC1), newspaper report (2008/6/LC1), job interview (2008/6/LC2), etc. Students may not have experienced similar settings so far, but these scenarios are what they are highly likely to encounter in future. After graduation, they may be interviewed in English in their job hunting; their occupations may require good command of English in communication; even in their vacation or business trips to foreign countries they need to use English to handle hotel booking and meal ordering.

To sum up, first, the interpretations we make about students’ listening ability are meaningful to the course syllabus, since both short and long conversations are quite congruent with CECR, stating that students should be able to follow classroom instructions, everyday conversations, and lectures on general topics conducted in English. Second, according to Bachman and Palmer (2010), a target language use (TLU) domain is defined as a specific setting outside of a test itself that requires the test taker to perform language use tasks. After examining both short and long conversations, we find that the settings of conversations share great similarity with the real life domains and these test tasks can be described as TLU tasks. In other words, these two test tasks bear high degree of authenticity by simulating the characteristics of target-language use in the real world. The interpretations we make about students’ listening ability are therefore generalizable to tasks of the TLU domain.

The following part will continue to describe the topic distribution of other test tasks in listening and reading components. Based on the classified topic areas mentioned earlier, the passages were grouped into three categories. Table 7.21 presents topic domains of short listening passages and compound dictation. It can be seen that social science is the largest category in short listening passages, and fourteen out of twenty-one passages fall into this category. For compound dictation, social science and humanities share equal proportion, and only one out of seven passages is natural science topic.

Table 7.21 Topics of short listening passages and compound dictation

Test paper (2006/12-2009/12)	Humanities	Social sciences	Natural science
Short listening passages	2	14	5
Compound dictation	3	3	1
Total	5 (17.9%)	17(60.7%)	6 (21.4%)

With regard to reading comprehension, Table 7.22 shows that seven fast reading passages all belong to social sciences. For banked cloze, humanities and social sciences take up the largest proportion, and only one passage belongs to natural science. In terms of careful reading, social sciences still ranks the top, followed by natural science and humanities.

Table 7.22 Topics of banked cloze, fast and careful reading passages

Test paper (2006/12-2009/12)	Humanities	Social sciences	Natural science
Fast reading		7	
Banked cloze	3	3	1
Careful reading	1	8	5
Total	4 (14.3%)	18(64.3%)	6 (21.4%)

This part did not make a diachronic comparison on the topic coverage of the pre- and post-2006 CET-4, since the classification of topic domain changed with the revision of NCETS. However, the two tables can still help summarize some characteristics from the above topic distributions.

First of all, topics of CET-4 passages on the whole are varied covering each category, which goes well with syllabus requirements that selected materials should be on general topics and do not favor a particular major of students. These topics range from education to economy, from environmental issues to medical science, from internet to modern technology. According to Weir (1993, p.67), topic should not be culturally biased or favor one section of the test population. No passage in the CET-4 involves topics like religion and politics, or a specific field of background knowledge. More topics are concerned with what human beings are universally confronted with such as global warming, female occupation, peace and war, or daily issues like how to be an energetic person, children's education, pressure.

Second, in both tables social sciences tops the rest, accounting for 60.7% in listening passages and 64.3% in reading passages. As mentioned earlier, social science as an umbrella term covers extensive aspects. There are passages whose themes test takers are quite familiar with like online learning, university scholarship, and themes that students are not frequently exposed to like customers' satisfaction, security of privacy, etc. From test designers' perspective, social science domain provides a broad selection of input materials so that culture and background bias can be effectively reduced and students' language ability can be assessed more objectively and thoroughly. Meanwhile, passages with familiar themes are interesting and authentic for test takers, while those unfamiliar ones can be informative and enlightening. In addition, from a broader perspective, this can also exert beneficial washback on college English teaching and learning. Textbook compliers and publishing house will include a variety of teaching materials. Students will expose themselves to more extensive topic domains to expand their horizon and enrich their topical knowledge.

Natural science takes the second place, sharing 21.4% in listening and reading passages respectively, involving passages related to pain management, male health, tracing criminals by hair, etc. The slightest distribution is occupied by humanities, totally accounting for 17.9% in listening passages and 14.3% in reading passages, covering themes like writing methods, importance of books, children's language, etc. According to Alderson (2000, p. 62), in the whole, non-specialist texts in the arts and humanities, and to some extent in the social sciences, will be easier to process for more people of equivalent educational background than scientific texts. This may help justify such topic distributions. If the topics are too unfamiliar to them, test takers may fail to get the details in spite of their grasp of the main idea. It follows that their test performance will be affected and their real abilities cannot be measured accurately, which indicates that the validity of the tasks is not as desirable as test designers expect. Since natural science texts are comparatively more difficult, a fair proportion of these passages contribute to discriminating students with higher language proficiency. Texts of humanities, being easier to handle, are included but their proportion is minimized especially in careful reading passages. Social science is controlled as the largest topic category with more passages included in this domain. However, on the other hand, even though the overall topic distribution is rational and can be justified, there is room for improvement. Take fast reading for example, seven passages are all from domain of social science, which may make students more focus on the structure and language characteristics of social science while ignoring other topic domains when they choose fast reading materials for practice. It is advised that CET-4 designers make tiny adjustments so as to make topic distribution of fast reading more balanced.

Third, the 2006 CET-4 syllabus also stipulates that passages should be selected from original materials for native speakers, including daily conversations, lectures, radio and television programs, newspapers, magazines and academic journals. Several reading passages were traced to their origins as evidence. In test paper of Jun. 2007, the first passage about identity fraud was originally from website of the US National Criminal Justice Reference Service (https://www.ncjrs.gov/spotlight/identity_theft/summary.html), and the second one on sex discrimination can be found on the website of the US Newsweek, which was published on December. 18, 2006 (http://www.newsweek.com/id/44140). The second careful reading passage of 2008/6 is about privacy security, which was adapted from a report published on October. 17, 2006 in American MSNBC website (http://www. Msnbc.msn.com/id/15221095). Just as analyzed previously in both short and long conversations, reading materials largely embody characteristics of authenticity.

7.3.2 Characteristics of expected response

The response is complex in that it should be distinct between the expected response and the test takers’ actual response (Cohen, 1980). The second part of this chapter, will discuss the expected response of CET-4 listening and reading components, which as another facet of test method, affects performance on language tests along with input.

7.3.2.1 Listening skills coverage

Table 7.23 presents listening skills specified by both teaching and testing syllabuses. According to Buck (2001), these taxonomies of sub-skills do help us think about what processes we should include in listening tests. Compared with the general stipulation of grasping main ideas and key points in the CECR, the CET-4 Syllabus provides an operationalized definition of listening construct with a detailed taxonomy of seven specific listening skills.

Table 7.23 Listening skills required by teaching and testing syllabuses

Syllabus	Listening skills
The Requirenents (2004)	Students are expected to employ basic listening strategies to facilitate comprehension. They should be able to grasp main ideas and key points.
The CET-4 Syllabus (2006)	A. Understanding main idea and important details 01 understanding gist 02 understanding important and specific details 03 determining speaker's opinions and attitudes B. Understanding inferences 04 making inferences and deductions 05 recognizing communicative functions of utterances C. Understanding meaning through linguistic features 06 recognizing phonological features (stress, intonation, etc) 07 determining semantic relationships such as comparison, cause, result, degree, purpose, etc

Based on Weir's (1993) taxonomy of communicative listening sub-skills, the first three skills (01-03) in category A require direct meaning comprehension; skills (04-05) in category B require inferred meaning comprehension; and skills (06-07) in category C are related to contributory meaning comprehension. Table 7.24 presents the skills coverage of the three test tasks in listening comprehension.

Table 7.24 Listening skills coverage of three task types

Test paper (2006/12-2009/12)	01	02	03	04	05	06	07	Total
SC		17	5	34				56
LC	4	39	1	4			1	49
SP	3	59		6			2	70
Total	7	115	6	44			3	175
	4%	65.7%	3.4%	25.1%			1.7%	

Note. SC=short conversations, LC=long conversations, SP= short listening passages

In terms of the short conversations, the most tested skill is *making inferences and deductions* (04), followed by skill *understanding important and specific details* (02). *Determining speaker's opinions and attitudes* (03) is examined by five items. Since these one-turn conversations are only small chunks conveying limited amount of information, the majority of items are designed to test students' inferential ability. Therefore, this task has a higher demand for students' inferred meaning comprehension

In contrast, long conversations present different skill distributions. The skill *understanding important and specific details* (02) becomes the most tested one, with thirty- nine out of forty-nine items checking it, while the skills *understanding gist* (01) and *making inferences and deductions* (04) come to the second. Since long conversations have an average of eight turns, it is rational that direct comprehending of main idea and details is more examined. Among the forty-nine items, only one item requires test takers to *determine speaker's opinions* (03), and one item to check their *semantic knowledge* (07).

Short listening passages share similar features with long conversations in that the skill *understanding important and specific details* (02) makes up the largest proportion, followed by the skill *making inferences and deductions* (04). We can see from Table 7.24 that fifty-nine out of seventy items are intended to check students' ability to *get specific details* (02). Only three items require *understanding gist* (01), and two items check knowledge on *semantic relationships* (07).

Taking an overall look at Table 7.24, we can draw some conclusions on skills coverage in listening comprehension. First, most of the listening skills stipulated in the test syllabus have been covered in the seven test papers except skills *recognizing communicative functions of utterances* (05) and *recognizing phonological features (stress, intonation, etc)* (06), which are not examined by a particular item. The possible reason is that *recognizing communicative functions of utterances and phonological features (stress, intonation, etc)* is most fundamental to understand listening input. Although no items directly check these skills, they are essential for listening comprehension and are actually examined along with other skills.

Secondly, different listening skills are examined by different test tasks. The skill *making inferences and deductions* (04) occupies the largest proportion in short conversations but takes the second place in long conversations and short listening passages. On the other hand, long conversations and listening passages attach more importance to checking students' ability to *understanding important and specific details* (02), while inferential and deductive ability come to the second in these two tasks.

Finally, if examining the total number, we can find that 115 out of 175 items (65.7%) are designed to check the skill *understanding important and specific details* (02), while inferential skill takes the second largest proportion (25.1%). Skills labeled as 01, 03, and 07 take up slight proportion, altogether no more than 10%. Overall, direct meaning comprehension is highly required and frequently examined in CET-4 listening tasks.

7.3.2.2 Reading skills coverage

Similarly, Table 7.25 presents reading skills required by teaching and testing syllabuses. The CET-4 syllabus lists three categories consisting of nine reading skills. The first two categories (A & B) cover skills employed in careful reading. The defining feature of careful reading is that the reader attempts to handle the majority of information to establish accurate comprehension of explicitly stated main ideas and supporting details, make propositional inferences, indentify lexis and understand syntax (Urquhart & Weir, 1998; Weir, 2005). Category C involves expeditious reading skills. Skimming means reading for gist, while scanning refers to reading selectively, involving locating specific information such as figures, names, dates of particular events.

Table 7.25 Reading skills required by teaching and testing syllabuses

Syllabus	Reading skills
Requirenents (2004)	Students are expected to employ effective reading strategies. They should be able to grasp main ideas and understand major facts and relevant details.
CET-4 syllabus (2006)	A. Distinguishing and understanding main idea and important details 01 understanding explicitly stated conceptions or details 02 understanding implicitly stated conceptions or details (e.g. conclusion, judgment, inference); understanding texts through semantic communicative functions(e.g. plead, refusal, command, etc) 03 understanding main idea (e.g. locating key points) 04 understanding the author's opinions and attitudes B. Employing linguistic skills 05 understanding vocabulary (e.g. deducing meaning of words or idioms from contexts) 06 understanding semantic relations (e.g. causes, result, purpose, comparison, etc) 07 understanding discourse (through lexical and grammatical cohesion devices) C. Employing specific reading skills 08 skimming for gaining main idea 09 scanning for locating specific information

Based on the nine skills, the analysis was conducted on fast reading and careful reading passages. Banked cloze is not included here because its task format mainly checks test takers' lexical and discourse knowledge.

Fast reading

The skills *skimming* (08) and *scanning* (09) are mainly examined in fast reading part. As Table 7.26 displays, sixty-nine out of the seventy items are designed to evaluate students' ability to locate specific information such as figures, names, etc, with only one item targeting at their ability to grasp main idea. It can be seen that the scanning skill has been fully examined in the past seven test papers, but the skill *skimming* (08) seems to have been ignored. However, it is noteworthy that all the fast reading passages are presented with a title, which serves as a quick prompt to facilitate test takers to predict what they are going to read and as a clear guidance for them to follow the major theme in their reading. Since a title does facilitate for grasping main idea, it may explain why few items covering the skill *skimming for gaining main idea (08)* were designed.

Table 7.26 Reading skills coverage in fast reading and careful reading passages

Test paper	01	02	03	04	05	06	07	08	09	Total
Fast reading								1	69	70
Careful reading	18	41	5	2	1	2	1			70
	25.7%	58.6%			15.7%					

Careful reading

As mentioned earlier, category A and B encompass seven careful reading skills listed in the CET-4 Syllabus. A general look at Table 7.26 shows that all the skills can find their places in spite of different proportions. However, even though the CECR stresses that students should be able to grasp main ideas and understand major facts and relevant details, large proportions of test items are designed to examine students' ability to *understand explicitly and implicitly stated* *conceptions or details* (01-02) while ignoring the skill *understanding main idea* (03).

The skill *understanding implicitly stated conceptions or details* (02) takes the largest percentage (58.6%). Such items usually require test takers to read between the lines and to infer from context. It poses higher demands for students' reading ability. The skill *understanding explicitly stated conceptions or details* (01) takes the second place (25.7%). However, these items are not superficially explicit so that test takers can immediately choose the answer by simple matching the same words in the passage with those in the options. They still need to paraphrase one or several sentences to figure out the correct option.

The rest of skills (03-07) only take a slight percentage (15.7%). Among them the skill *understanding main idea* (03) is particularly stressed in CECR. The skill *understanding the author's opinions and attitudes* (04) are demanding for test takers' proficiency. Nevertheless, examinations on these two skills are largely ignored. Hence, it is suggested that these skills should be strengthened by increasing the number of corresponding items.

In terms of skills 05-07, they require lexical, semantic and discourse knowledge. Few items are designed directly to examine test takers' employment of linguistic skills stipulated in CET-4 syllabus, given that linguistic knowledge serves as a basis for test takers' comprehension and facilitates their accurate understanding of reading materials. That may be the possible reason why these three skills occupy the smallest proportion.

7.4 Summary

In this chapter, two sources of backing were presented first to examine construct of the CET-4. A large pool of students' CET-4 test scores underwent statistical analyses. The results proved that listening, reading, integrative and writing skills were all measured by the test paper, which in turn all contributed to one general factor. In order to make the score-based interpretations meaningful, the construct definition of the CET-4 was explored through document analyses. Established on a combination of the uniform teaching and testing syllabuses, and a current language ability model, the construct to be measured by the CET-4 is defined as the overall English proficiency, to be more specific, Communicative Language Ability in technical terms. Evidence from these two sources show that CET-4 scores can be interpreted as indicators of test takers' language abilities. The score-based interpretations can provide meaningful information and relevant to decisions to be made.

The larger part of this chapter reported on the major findings of content analysis of listening and reading components. The analyses were conducted from characteristics of input and characteristics of expected response, including five parameters: text length, readability, genres, topics, and skills coverage. The results were analyzed based on the CECR and the CET-4 Syllabus, both offering basic and authoritative criteria for evaluation. Findings were further discussed with reference to a few previous theoretical and empirical studies. The analyses of seven test papers have proven that the revised listening and reading components in the CET-4 overall possess a higher degree of content validity. Major conclusions can be summarized as follows:

First, text length of short listening passages ranges from 217 to 284, and for compound dictations, it is between 173 and 237. The number of turns in long conversations ranges between five and twelve. Text length of listening and reading passages, in the whole, are quite congruent with requirements stipulated in both CECR and 2006 CET-4 syllabus.

Second, the average readability scores for short listening passages (61.0) and compound dictation (59.6) fall into standard scale. The average readability scores for banked cloze (56.1) and careful reading (53.8) fall into fairly difficulty scale while readability of fast reading (47.8) into difficult scale. The figures reveal that test difficulty of listening components, as the two syllabuses state, is less difficult than that of reading components. However, according to the two syllabuses, the difficulty level of fast reading should have been less difficult. Hence, it is advised that the difficulty level of fast reading should be adjusted and held consistent.

Third, each genre specified in the two syllabuses is representative with certain proportions of passages. Narration is the mostly used genre in listening passages while argumentation is the least. Exposition and argumentation are two major genres adopted in reading passages. The genre distributions in listening and reading components can be justified as reasonable and scientific.

With regard to topics, its range is extensive and diverse. Topics of short conversations are related to students' daily life and academic study, to which students are frequently exposed. Topics of long conversations involve scenarios of job hunting, interviews, hotel checking, etc., in which students are highly likely to engage in their future work and life. Hence, it can be concluded that conversation parts demonstrate higher degree of authenticity and great similarities with TLU domains. In terms of listening and reading passages, generally speaking, social science is the mostly adopted topic area, followed by natural science and humanities. One point deserves attention is that all the seven fast reading passages were chosen from topic area of social science, so adjustment is advised to make its topics more balanced and diversified.

Characteristics of expected response were examined by checking skills coverage in listening and reading tasks. Almost all the skills listed in the CECR and the CET-4 Syllabus have been covered in the seven test papers, with different skills predominantly checked by different tasks. Short conversations lay more emphasis on examining students' skill *making inferences and deductions* (04). The skill *understanding important and specific details* (02) is attached great importance in both long conversations and short listening passages. As to reading components, the skill *scanning for locating specific information* (09) is fully examined in fast reading while skills *understanding both explicitly and implicitly stated conceptions or details* (01-02) are frequently tested in careful reading.

In conclusion, in spite of some small discrepancies or inconsistencies, statistics from content analysis demonstrate an overall neat correspondence in the aspects of the six parameters between what the test papers are intended to test with what is stipulated in both teaching and testing syllabuses. Findings from the above discussions reveal that the content validity of the revised listening and reading components since the 2006 CET-4 reform is acceptable.

CHAPTER 8

BACKING FOR THE DECISION CLAIM

8.1 Introduction

Claim 2: The multiple decisions that are made on the interpretation of the CET-4 scores reflect the existing educational and societal values and the relevant university regulations, and are equitable for all the stakeholders to be affected by the decisions.

This chapter reports on backing evidence for the decision claim which is presented in the above box and its corresponding research question (RQ2): what evidence has been provided or is needed to justify the major types of decisions made based on CET-4 scores. Results from interviews, questionnaires, and document analyses are discussed in support of the decision claim. The chapter first summarizes the specific decisions that are made on CET-4 scores, and zooms in on the corresponding decision makers and stakeholders to be affected by these decisions. It continues to discuss factors underlying these decisions from perspectives of values and equitability. Finally, the chapter suggests some backing evidence that universities should provide to be held accountable for stakeholders to be affected, and draws attention to potential rebuttals threatening the legitimacy of the decision claim.

8.2 Major decisions made on CET-4 scores

In Chapter 4 Table 4.1 displays decisions made on CET-4 scores, stakeholders to be affected by the decisions, and individuals responsible for making these decisions (see section 4.4.3). The related CET-4 score-based decisions generally fall into three layers: nationwide decisions made by the NCETC, institutional decisions made by program administrators, and decisions made by employers at the social dimension. As stated earlier, the present study is situated within the instructional setting. Since using CET-4 scores as one of the criteria for employment is related to test use in social dimension, it will not be probed into in this study. Thus, this section summarizes the major types of decisions made by the NCETC and by universities sampled in the present study.

8.2.1 Nationwide decisions made by the NCETC

Since 2005, the NCETC has adopted a new score reporting system based on an overall score range from 220 to 710 with a mean of 500 and a standard deviation of 70. The Score Report Form (SRF) displays profile scores from each component and an overall score rather than a certificate. In spite of cancelling the certificate with a pass or distinctive, new cut-scores are established as qualification thresholds, which is a typical means to identify students' English proficiency and to facilitate understanding of students' achievements. Table 8.1 lists decisions directly made by the NCETC based on CET-4 scores. Test takers with scores beyond 220 will be awarded a SRF. Test takers with scores beyond 425 are qualified for taking the CET-6, and those with scores beyond 550 are entitled to take the CET-4 SET.

Table 8.1 Nationwide decisions made by the NCETC

Multiple decisions	**Stakeholders to be affected by the decisions**	**Corresponding decision makers**
Set different CET-4 cut-off scores ● 550 for taking CET-4 SET ● 425 for taking CET-6 ● 220 for issuing the Score Report Form	students teachers the University Academic Affairs Office	the NCETC

As discussed earlier, one of the major criticisms on the CET-4 stemmed from misusing its certificate as one of the prerequisites for graduation or employment decision. The decision made by the NCETC to abandon issuing test takers the

CET-4 certificate is thus to avoid it continuously being overemphasized. Reforming the score reporting system is a measure to mitigate negative washback of the CET-4. Just as Jin (2006, p.8) indicates, it is more used as “a purposeful approach to encourage the use of CET for EFL teaching and learning purposes and avoid putting undue pressure on the test and its designers”. This approach to some extent has achieved the intended purpose of encouraging rational use of test scores. According to a study conducted by Wang and Wang (2012), among the 540 nationwide universities they investigated, 63% of them have no longer linked CET-4 performances to academic degrees. About 140 universities still insisting on this requirement have loosened this policy by flexibly setting the cut-off line from 330 to 470 based on the teaching and learning situations in their own universities.

A full evaluation of the score reporting system should include consideration of major stakeholders’ perceptions of it, so both students and teachers were asked to evaluate whether the CET-4 SRF can better reflect students’ English proficiency than the traditional certificate. From Figure 8.1 it can be seen that comparatively speaking teachers were more positive of the current Score Report Form, since over half of the teacher respondents agreed (43%) and strongly agreed (21.1%) with the statement while only about one third of students surveyed agreed (27.4%) and strongly agreed (8.0%) with it.

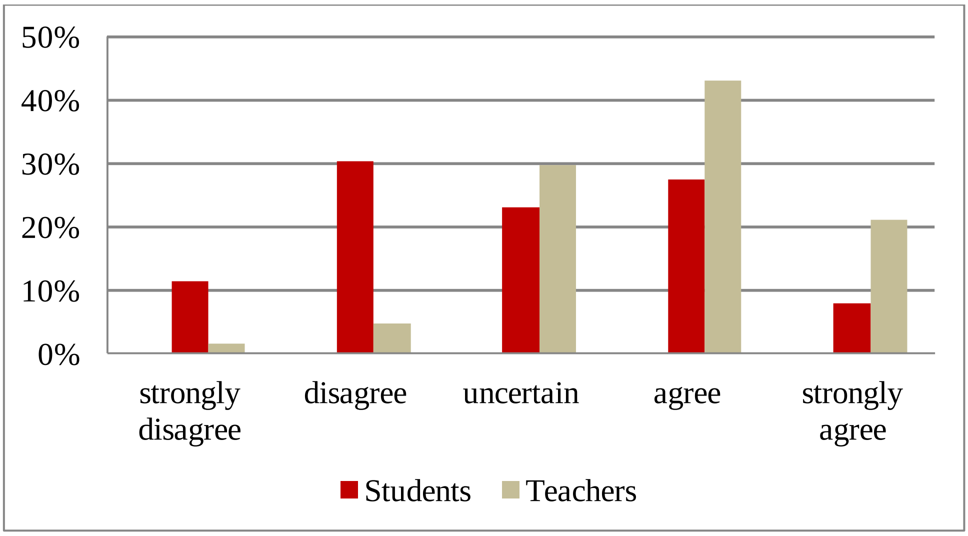

Figure 8.1 Students’ and teachers’ attitudes to the SRF

It seemed that a discrepancy between students' and teachers' attitudes towards the Score Report Form emerged. Thus an independent t-test was further performed. Table 8.2 shows that 753 students had a mean of 2.9 and 128 teachers had a mean of 3.77 on a 5-point Likert scale of agreement. The t-test revealed that their means differed significantly at the p < .05 level, which also confirmed that teachers held positive perceptions of the SRF, while students held less positive attitudes than their teachers did.

Table 8.2 Independent-sample t-test of students' and teachers' attitudes to the SRF

Variable	Group	N	Mean	SD	T-value	df	Sig. (2-tailed)
SRF	Students	753	2.90	1.159	-8.109	879	.000
	Teachers	128	3.77	.889			

Since this attitudinal discrepancy could not be explained by questionnaire findings, this question was further brought about in group interviews, some students commented that:

> The Score Report Form can help us diagnose our strengths and weaknesses in particular of the four language skills. However, we feel a little bit confused at the 220-710 score scale since we were tested by a 100-point test paper. What we are quite clear is the cut-off line 425 means we can get our Bachelor's degrees and we can also take the CET-6.
>
> (The first student group interview, June 6, 2010)

> With a 100 points score scale, my classmate may score several points higher than me, while our performances may have a difference of dozens of points with adoption of 210-710 score scale. However, the potential employers do not know the score calculation and may assume our language proficiency varies a lot.
>
> (The first student group interview, June 6, 2010)

Teachers' positive attitude to some degree may be due to their basic understanding of the score transformation process described by the NCETC:

> We can better understand advantages of norm-referenced score calculations. The profile scores to each component obviously provide more information about students' English proficiency. We welcome this modification.
>
> (The first teacher group interview, June 11, 2010)

Considering one of the major reasons to change the score reporting system is to reduce the social weight put on the test, it is interesting to notice that some interviewees expressed unexpected voice:

> Compared with the traditional certificate only marking pass or failure, the current SRF actually exerts more pressure on us. In job hunting market, the potential employers can immediately evaluate our English proficiency by score differences.
>
> (The third student group interview, June 7, 2010)

In spite of bringing about heavy pressure, the SRF may discourage students from aiming at a narrow pass of 425 points. Just as one interviewee said:

> We need to make more efforts to improve scores if we want to impress employers and gain career edge with excellent English proficiency. In addition, we need to improve the four language skills given that profile scores from components are all displayed on the SRF. You never know which skill is more demanded or valued in your future workplace or by your potential employers.
>
> (The third student group interview, June 7, 2010)

The interviews indicated that students were aware that inclusion of profile scores could reflect their English proficiency more accurately as well as better diagnose their strengths and weaknesses. Their less positive attitude to the SRF was in part due to the pressure that they should score as high as possible to impress employers in the competitive job hunting market.

With regard to the cut score for taking the CET-4 SET, strong dissatisfaction can be perceived:

> The cut-off line (550) is so high that only a few students in my class have qualifications to take it. Most of us have no opportunity to check our oral English. We do not know why they set such a higher cut score.
>
> (The second student group interview, June 6, 2010)

Another respondent echoed the discontent:

> It is unfair because my oral English is better than my reading ability. If I get the chance to take the oral test, I am able to get a desirable grade.
>
> (The second student group interview, June 6, 2010)

Questionnaire respondents expressed similar opinion. When asked about whether the CET-4 SET should be open to all the students, both students and teachers showed approving attitudes with the mean values of 3.57 and 3.84 (see Table 8.3). When they were further asked about whether students would spend more time improving their oral English proficiency supposing the SET was compulsory, the mean value of agreement reached around 4.0.

Table 8.3 Students' and teachers' attitudes to the CET-4 SET

	Students		Teachers	
	Mean	SD	Mean	SD
The CET-4 SET should be open to all the students.	3.57	1.099	3.84	.978
If the CET-4 SET is compulsory, students will spend more time improving their oral English proficiency.	3.96	0.821	4.06	0.801

Whether the SET should be open to all test takers has been a long-standing issue. The practical constraint is that the face-to-face SET involves high costs and more resources like raters, rater training, video recording, etc. Since its trial operation in 1999, the NCETC has establish 51 oral testing centers in 36 provinces and the annual SET candidature has reached around 0.1 million (Jin, 2005). However, compared with 10 million test population for the written CET, the number of oral test takers actually occupied a small fraction. Thus, it is reasonable to doubt whether the CET-4 can maximize its positive washback by pushing students to spend time cultivating their oral English proficiency, and whether the CECR can realize its objective to improve students' language use ability especially listening and speaking abilities. In response to stakeholders' discontent, the NCETC declares that given the successful administration of semi-direct oral proficiency test in some large-scale tests, they have conducted an ongoing research project on computerized CET SET in hope of enlarging its scale to accept more test takers and maximizing its positive washback (Jin, 2005; Jin & Guo, 2002).

8.2.2 Institutional decisions made by the University Academic Affairs Offices

Not all the universities show quick and positive responses to the NCETC's reform in the score reporting system. Some universities still keep their institutional practices of setting the cut score of 425 as one of the prerequisites of conferring the Bachelor's Degree. In addition, more additional uses are added to CET-4 scores. This section delineates score-based decisions at institutional level based on results from questionnaires and interviews. Table 8.4 lists major types of decisions identified in the four sampled universities in this study. Even though they are kind of case studies in nature, it reveals typical practices in quite a large number of universities. These institutional decisions are expected to be generalizable to and representative of common practices in university community of Xi'an.

First, graduation decision is made on CET-4 scores in the four universities. According to the student questionnaire, 87.5% of surveyed students admitted that their universities imposed such a requirement that students with CET-4 scores below 425 cannot be conferred their Bachelor's degrees. Administrators from the University Academic Affairs Offices also confirmed this decision in interviews.

Second, placement decision is made on CET-4 scores in U1. The cut score (425) for students to take the CET-6 is used as a threshold to place students into CET-4 or CET-6 preparation courses which are usually open in the middle of the second semester of Year 2.

Table 8.4 Institutional decisions made by University Academic Affairs Offices

Multiple decisions	**Stakeholders to be affected by the decisions**	**Corresponding decision makers**
Graduation decision (set the CET-4 score 425 as the threshold to award students BA (U1,U2,U3,U4)	students	the University Academic Affairs Office
Placement decision for CET-4 and CET 6 preparation course (U1)	students	the University Academic Affairs Office
Pass/fail in EFL course (U1,U2)	students	the University Academic Affairs Office
Adjust teaching curriculums based on students' CET-4 performances and passing rate (U1,U2,U3,U4)	teachers students	the University Academic Affairs Office

Third, both U1 and U2 use students' CET-4 overall scores as their score records of English course in the second semester of Year 2, replacing the usual university-administered final English examination.

Fourth, decisions to adjust teaching curriculums are made on the overall CET-4 performances or passing rate of their university students. Three universities (U1, U3, U4) all have CET-4 preparation courses. In U4 all the students are required to take the CET-4 at the end of Year 1, and in view of their almost 100% of accumulated CET-4 passing rate, a wide range of English elective courses are open for students to choose from during Year 2. In U2 the class periods of the listening course have been increased from two to four per week, given that the current CET-4 is more focused on measuring students' listening ability. With the increasing CET-4 passing rate in recent years, the Academic Affairs Office in U2 also made the decision to abolish the traditional practice of opening CET-4 test preparation course.

The aforementioned discussions summarize major types of decisions made on CET-4 scores, which evidence again that multiple decisions tend to be made on scores of the same assessment. However, the argument is whether these decisions can be justified with appropriate and convincing evidence to support its uses. To make sure these decisions can generate positive consequences, qualities of value-sensitivity and equitability should be taken into account in the decision-making process.

8.3 Warrants on value-sensitivity of decisions

Warrant A1: Decisions made on CET-4 scores take into account the existing educational and societal values against the background of Chinese testing setting.

Evidence: No documents or reports from the NCETC or the universities can be traced to address the decisions from perspectives of educational and societal values. Lack of evidence in this aspect may become a rebuttal to challenge the qualities of the decision claim.

Discussion: Just as Bachman (2005, p.29) stressed, “Test development and use always take place in a value-laden sociopolitical context”, this study thus explores China’s societal and cultural values embedded in test use. The history of examination in China can be traced back to Han dynasty, with tests mainly used to select officials (Cheng, 2010; Spolsky, 1995). This imperial examination system strengthened the utilitarian values of education and the role of examination in changing one’s life (Han & Yang, 2001). Today the Chinese educational culture is still characterized as an examination-oriented system, in which testing and examinations remain their important and powerful role (Cheng, 2008, 2009; Li, 1990; Qi, 2005). Using tests as a gateway to selection, advancement, or competition, and acknowledging examinations as a tool ensuring fairness take a deeper root in the inherited influences of the imperial examination system.

The implementation of the CET-4 serves as a tool to check university students' English proficiency. The cut scores set by the NCETC facilitate to distinguish students with varied language proficiency, and allot testing resources to accommodate test takers of CET-6 and SET. In addition, it also provides benchmarks for the potential employers in making recruitment decisions, even though the use for employment is not what the test is intended for.

These decisions all implicitly embody the traditional value that examinations have been accepted as a comparatively fair tool for selection. In China, students have to take numerous examinations throughout their schooling. After stepping into society, they may still need to take examinations for certification and career development. It is thus the commonly held societal belief that higher test scores tend to bring a person more opportunities, advance him to a higher educational or professional level, and even bring success to one's life. These inherited and deep-rooted educational and societal values have played an indispensable role in nurturing the increasing power of a test and the high-stakes of score-based decisions.

Institutional decisions made on CET-4 scores are also sensitive to educational values. Both the CECR and the CET-4 Syllabus specify that the major purpose of the CET-4 is to promote and positively impact college English teaching and learning. In spite of varied views on this function, it has to be admitted that modern education has witnessed a tendency of using large-scale and high-stakes tests as a catalyst or a lever for curriculum innovation (Andrews, 2004; Cheng, 2008; Jin, 2006; Qi, 2003). It explains why universities make institutional policies or adjust teaching curriculums in line with the test. As mentioned earlier, considering the Listening component in the CET-4 is modified with more score weight and varied task types, the Academic Affairs Office in U2 made the decision to increase the teaching hours of listening course from two to four class periods per week. In order to improve passing rate, U1, U3 and U4 all opened CET-4 preparation courses. It is evident that testing can exert either immediate or far-reaching influences on teaching and learning. Since the CET-4 is designed in line with the uniform teaching syllabus, it should serve college English teaching and learning. It is noteworthy for both test developers and test users to think about how to take advantage of the washback mechanism to maximize positive effects of this test and minimize its negative.

Warrant A2: Decisions made at the institutional level take into account the legal documents, relevant university regulations and common practices in the university community in Xi'an.

Evidence: No backing can be traced in official documents or institutional regulations to support the decision to link students' CET-4 scores to their Bachelor's degree. It thus exists as a hidden policy. Without providing convincing backing to justify these institutional decisions, universities are easily subject to criticisms and even lawsuits against these decisions. The most challenging rebuttal to the institutional decision from test takers is targeted at the graduation decision, which is articulated below:

Rebuttal: Linking students' CET-4 performances to their academic degrees violates related legal documents and university regulations.

Discussion: In recent years, there have been news reports about students' lawsuits against their universities for not conferring them Bachelor's degree only due to their failure to pass the cut-off line set by universities. Students accused that linking graduation decision to their CET-4 performances had no legal basis, and listed the following rebuttal backing to support their lawsuits.

Rebuttal evidence: First, no article can be found in the Higher Education Law of the People's Republic of China stipulating that CET-4 performances can be used as one of the criterion to confer academic degrees. The Article 20 related to granting degrees was quoted as supporting evidence:

> Students receiving higher curricula education shall be issued corresponding certificates of educational background or other certificates of studies by the institutions of higher learning or scientific research institutes approved to undertake the task of post graduate education they have been in on the basis of the length of schooling and achievements in studies in accordance with the relevant provisions of the state.
>
> (Ministry of Education, 1999)

Second, no provision in their universities' Student Handbooks spells out students must pass the CET-4 cut-off line set by their universities to get their Bachelor' degrees. Third, even though the university claimed that such a graduation decision had been an established convention, the cut score should be consistent rather than changeable. For instance, one university had set 320 points as the cut-off line for years but suddenly increased it to 328. Moreover, the student was informed of this change upon graduation so that he missed the opportunity to retake the CET-4 to secure a higher score. Students maintained that it was illegal to base graduation decision on the CET-4 scores given no any legal documents or university regulations in support of this graduation decision.

The legal cases discussed above reveal that it is of vital importance for universities to provide adequate and convincing backing to justify their decisions and to be held accountable for test takers.

Interviews conducted with administrators from the University Academic Affairs Offices also help interpret why the graduation decision is tied to CET-4 scores. Excerpts of their interviews are presented below.

> This requirement can ensure both students and teachers lay more emphasis on their English teaching and learning. In addition, the rank of our university has advanced in recent years, which in a large part should be attributed to the improvement of students' CET-4 performances, the higher passing rate. Thus, we cannot take risks to abolish this requirement.
>
> (Administrator of U1, personal communication, June 8, 2010)

> We insist on this hidden policy to discipline students to make more efforts to improve their English proficiency, since other universities in Xi'an all keep such a requirement rather than abolish it. However, considering the desirable passing rate in recent years, we no longer stop the normal teaching for test preparation at the approach of the CET-4 administration. Minimizing the influence of "teaching to the test" is our first step to reduce the negative influence of this test.
>
> (Administrator of U2, personal communication, June 11, 2010)

> The accumulated CET-4 passing rate before students' graduation can almost reach to 100% in our university. Hence, our students take the CET-4 at the end of Year 1, and are offered a wide range of English elective courses including literature, poetry, translation and newspaper reading, etc. in Year 2. We still keep this policy since it does not induce strong protests or criticisms from our students. We assume this requirement plays a more stimulating role in our students' learning.
>
> (Administrator of U4, personal communication, June 10, 2010)

It can be seen that administrators from the three universities' Academic Affairs offices expressed similar reasons for linking CET-4 scores with gradation decision. They know some universities in major cities like Beijing, Shanghai have either cancelled this policy or loosened it by flexibly setting a lower cut score. However, considering all the universities in Xi'an still hold onto this policy, they prefer to keep this common practice as well. They expect that the high-stakes of the test injected by the graduation decision can be used as an impetus to stimulate students to take college English study seriously.

Some of the above comments were evidenced by questionnaire findings. In response to the graduation decision, half of the surveyed students agreed (42.6%) or strongly agreed (9.3%) that linking CET-4 scores to their Bachelor's degree could motivate their English learning, with 17.4% of them being uncertain and 30.7% showing negative attitude. It revealed that students also held mixed feelings.

To sum up, test influence is like a double-edged sword. In China, it is assumed that higher test scores can represent better achievements of students, and better achievements of students can reflect a university's higher achievements in teaching effects and administration, which in turn may earn a university wider fame and higher rank. If the related educational departments evaluate universities by taking into account their students' CET-4 passing rate, universities are likely to push students to improve their test performances by making certain high-stakes decisions such as linking it to their academic degrees. This also serves as a case in point on how a test is used as a powerful tool in the centralized educational system. It deserves consideration whether it is ethical for students to take the stakes for universities' fame and rank.

The argument this section wants to highlight is that decision makers must provide backing to whatever decisions they make. It is a little bit weak to claim that the graduation decision is made to follow the common practices of university community in Xi'an, since it is easy to be refuted by challenging why not follow universities in Shanghai or Beijing to abolish this requirement. The graduation decision tends to induce ethical controversy if universities keep this hidden policy to guarantee higher CET-4 passing rate and ensure their ranks in nationwide university ranking list. Therefore, it is advisable for these universities to take some measures so that they can justify this decision on setting CET-4 cut scores as threshold of conferring academic degrees in the event of any dispute.

First, in response to students' citing the Article 20 in the Higher Education Law of the People's Republic of China as the rebuttal backing, it should be noted the Higher Education Law also acknowledges each university's autonomy and independent decision-making rights to run university's educational and administrative system, on the condition of abiding by laws and regulations related to higher education. Since the Higher Education Law is applicable to all the institutions of higher learning, articles in it tend to be generalizable. Therefore, it is reasonable and legally permitted for each university to issue its own institutional policies by adding locally specific provisions or crystallizing provisions pertaining to granting academic degrees. In addition, this requirement should not exist as a hidden policy. Universities are advised to incorporate it into Student Handbook and related institutional regulations so that it can be traced as evidence in case of any lawsuits. Above all, it should be guaranteed that students and other major stakeholders are informed of this requirement far ahead of taking the test, which is to be discussed in the warrant on equitability.

Second, when universities made an argument that this graduation decision served as an impetus to motivate them to improve English, they need to conduct research exploring students' perceptions of the CET-4 and of institutional policies related to it, test influences on their learning motivations and behaviors, etc. This is also in accordance with suggestions proposed by Bachman and Palmer (2010) that in the decision-making process stakeholders should be widely consulted. In this way, universities may get evidence to convince students of the motivating power of this policy. For example, the questionnaire findings in this study show that over 50% of students admitted the CET-4 motivated them to study English hard. By the contrary, if students' feedback to this graduation decision is overwhelmingly negative, it at least can alert universities to take a serious concern of this policy, and come up with possible alternatives.

Furthermore, universities may set a constant cut-off line lower than score of 425 by taking into account the teaching objectives and students' English levels in their own universities, or a flexible cut-off line calculated on the basis of the overall CET-4 performances or passing rate of one cohort of test takers. Whatever the method to set a cut-off line, the way to make the decision should be transparent, informative and equitable to all the test takers, which is to be discussed in the next section.

8.4 Warrants on the equitability of decisions

Warrant B1: The same cut scores are used in making decisions and no other considerations are taken into account.

***Evidence*:** Given the far-reaching influence of a large-scale and high-stakes test, the interpretations of test scores should be reliable, valid, meaningful and fair to all the test takers. Cut scores are set to help identify test takers' proficiency and to make classification or selective decisions. It must follow accurate and scientific measurement procedures to avoid any classification errors. As to the CET-4, students' test performances are subject to score weighting, IRT equating, adjustment of writing scores, and score normalization. These statistical procedures have assured the CET scores are comparable and the score interpretation consistent across its different administrations. In addition, scores from every administration of the CET-4 are equated to keep the measurement criteria unchanged (Jin, 2006). Therefore, the uniform cut scores set by the NCETC have the same meaning for all the test takers no matter where and when they take the CET-4. Take the cut score of 425 for example, it means that students with scores beyond 425 have met the requirements set by the CECR for Band Four and they are entitled to take the CET-6. This score is applicable to all the nationwide university students regardless of universities' geographical locations, types and ranks.

Discussion: As discussed earlier, the four universities set the score of 425, which is originally the cut-off line set by the NCETC for taking the CET-6, as the threshold for students to get their Bachelor's degrees. The graduation decision is based on the same cut score, which is applicable to all the students in their universities regardless of their majors, performances in other courses or even performances in the university-administered English course examinations.

Warrant B2: Test takers, EFL teachers and other stakeholders within the university community are fully informed about how the decisions are made and whether decisions are actually made in the way described to them.

Evidence: In order to inform test takers, EFL teachers and other major groups of stakeholders of these cut scores, the NCETC has disseminated information related to scoring methods and the eligibility to take corresponding tests both on its official CET website and in the uniform CET-4 Syllabus. In addition, the website presents the total score and the table of percentile comparison of each component so that test takers can find their percentile positions and test users can learn about the English proficiency of a particular examinee.

For university's decision to set the CET-4 scores as a prerequisite for getting the Bachelor's degree, teachers are informed by the University Academic Affairs Office, and then tell their students about this decision. All the major groups of stakeholders know the existence of this decision but a written text of this regulation cannot be traced. That explains again why this is a hidden policy, enforced top-down by the university authorities.

It is advisable that universities inform students of institutional decisions made on CET-4 scores in advance and establish criteria for graduation as a written provision in university regulations and Student Handbook rather than a hidden policy. Moreover, the procedures and factors involved in the decision-making process should be explicitly established in order to make sure decisions are actually made in the way described to them.

Warrant B3: Test takers have equal opportunity to learn or acquire the ability to be assessed.

Evidence: The CET-4 is administered to check university students' English proficiency after they complete the two-year English study at the foundation stage. Some top universities permit their students to take the test at the end of first year because students admitted to these top universities are supposed to have higher language proficiency. During the second year these universities will open English elective courses such as newspaper reading, British and American literature or courses specific to their majors. Whether they are currently required to take the CET-4 or not by their universities, students are provided with equal opportunity to learn English since the CECR explicitly stipulates that College English, an integral part of higher learning, is a required basic course for undergraduate students (Ministry of Education, 2005, p.5). It thus follows the potential test takers have equal opportunity to learn or acquire the ability to be assessed.

8.5 Summary

This chapter investigated decisions that are made based on CET-4 scores by test developers (the NCETC) and test users (universities). Instruments of questionnaires and interviews were conducted with stakeholders of test takers, teachers and the University Academic Affairs Office to collect evidence in support of the decision claim. More focus was put on exploring factors of value and culture implications that play a role in making these decisions than on the washback of the test per se. It also identified backing that was lack of, pointed out potential rebuttals threatening the legitimacy of the decision claim, and proposed possible solutions to seek backing.

Tests are always intended to serve the needs of an educational system or of society at large (Bachman, 1990, p.279). The launch of the CET-4 is to promote college English teaching as well as provide a measure to check university students' English proficiency. Meanwhile it also meets social demands for graduates with higher English proficiency. The NCETC's decisions on test reforms reflect their awareness of state-of-art theories and practices in language testing community. The decisions made by the NCETC to set cut scores for students' taking higher levels of tests and the institutional decisions made by program administrators all take a deeper root in the inherited imperial examination system in China and long-standing societal values, that is, tests can be used as a gatekeeper for selection or classification purposes. The graduation decision is made since universities' expected scenario is that the high-stakes decisions would stimulate teachers and students to attach more importance to their English teaching and learning. The decision to adjust teaching curriculums reflect a trend in educational setting to use a large-scale and high-stakes test as a lever or catalyst to impact teaching and learning.

Messick (1995, p.748) stressed that, “What matters is not only whether the social consequences of test interpretation and use are positive or negative, but how the consequences came about and what determined them”. Since these consequences are brought about by decisions made on the assessment, the score-based decision inference, as an indispensable connection between interpretation of scores and consequences of test use deserves language testers’ attention. This study is more interested in investigating what motivated universities to make these decisions, exploring what consequences they expected resulting from the proposed test uses, and identifying what evidence is necessary to support each proposed test use. This study is expected to raise awareness of both test developers and test users that any decision made on test scores, particularly decisions beyond the test’ intended purposes should be supported by solid and convincing backing. Both test developers and test users should be held accountable to justify decisions they made to test takers and other affected stakeholders.

CHAPTER 9

BACKING FOR THE CONSEQUENCE CLAIM

9.1 Introduction

Claim 3: The consequences of using the CET-4 and of the multiple decisions that are made are beneficial to stakeholders.

This chapter reports on backing evidence for the consequence claim which is presented in the above box and its corresponding research question (RQ3): In what way and to what extent can the CET-4 and the decisions that are made on it affect English teaching and learning. This chapter comprises three sections in accordance with the three warrants under the consequence claim. The larger part of this chapter reports statistical analyses of questionnaires by addressing variables like stakeholders' perceptions of the revised CET-4, test preparation activities, teaching and learning behaviors. Findings from this part seek answers to how stakeholders perceive the CET-4 and its washback (RQ3.1). Inferential statistical analyses were also performed to explore how students' perceptions of the CET-4 affect their test performances (RQ3.2). In addition, qualitative analyses of stakeholders' interviews were conducted to complement quantitative findings from questionnaires.

9.2 Warrant on beneficence of test consequences

Warrant 1: The consequences of using the CET-4 are beneficial to immediate stakeholders including students, teachers, as well as universities.

Bachman and Palmer (2010) advise test designers to start with the claim of consequence in an AUA in test development, and specify the major groups of stakeholders in the first warrant. Detailed discussion of the quality of beneficence can be developed in the rest of warrants, particularly the warrant pertaining to washback. This study evaluates and justifies test uses, so a bottom-up approach is adopted. Stakeholders to be immediately affected and less directly affected have been discussed both in Chapter 4 (see section 4.4.4) and in the inferential link of the decision claim in Chapter 8 (see section 8.2.2). As noted earlier, students, teachers, and the University Academic Affairs Office have been identified as the three major groups of stakeholders. Thus, Warrant 3 will explore whether the consequences of the CET-4 are viewed as beneficial from perspectives of these three major groups of stakeholders.

9.3 Warrant on the Score Report Form

Warrant 2: The CET-4 Score Report Forms are treated confidentially, presented in clear and understandable ways, and released to test takers and the University Academic Affairs Office in time for them to be used for making decisions.

Evidence**:** This warrant in general can be supported by the CET-4 Syllabus, the articles published by the NCETC (Jin, 2006, 2008, 2010) and the information on the official CET website, which all describe and explain the reformed score reporting system and its Score Report Form. The NCETC declared that the SRF received positive comments from stakeholders (Jin, 2006).

Discussion: However, in addition to the evidence from document analysis, there is no direct and detailed evidence revealing how stakeholders think of the SRF. Thus, this section will discuss evidence provided by the NCETC as well as findings emerged from the present study.

Confidentiality of SRF

According to Bachman and Palmer (2010), keeping assessment reports confidential is an issue related to fairness. In order to protect test takers' fundamental rights, their assessment records should be provided to test takers themselves and individuals who are authorized to receive them. For the CET-4, three ways are offered for test takers to get access to their scores. The most popular way is to check scores on internet. Currently, two websites (http://www.chsi.com.cn/, http://cet.99sushe.com/) are authorized by Ministry of Education and the NCETC to report CET scores. The former only requires a test taker to input his name and candidate number, while the latter requires one more procedure of downloading and installing an encryption key first for the sake of security. After a test taker logs in, his name, affiliated university and test score profiles will be displayed on the webpage. An alternative way is to text a designated number and a test taker's candidate number to China Mobile Limited, China Unicom, or China Telecom, which costs one RMB, and then a message displaying test scores will be sent back. The most prolonged way is to get the official Score Report Form from universities, because the printing of SFR and its delivery from the NCETC to universities take quite a period of time.

However, there is still a small chance that anyone knowing a test taker's test number can easily know his scores either by logging on the websites or texting a message via a mobile phone. The possibility of tests scores being leaked or misused depends on whether test takers can keep their candidate numbers confidential. It seems that CET-4 developers and the NCETC skillfully shift their responsibilities of keeping scores secure to test takers. We can examine what ETS and Cambridge ESOL have done in this regard. TOEFL or IELTS for example, a test taker for TOEFL or IELTS is advised to register online and create a personal profile with his username and password. He must log on his online account before getting access to his scores. Comparatively speaking, it is securer and more advisable for a test taker to be informed of test scores via his personal online account.

Clarity and understandability of SRF

Since June 2005, the NCETC has implemented a new score reporting system, with an overall score range of 220 to 710, a mean of 500 and a standard deviation of 70. There is no pass or fail. Test takers with total scores beyond 220 will be issued a Score Report Form with both the overall score and the profile scores for each component: the overall score (710 points, 100%), listening comprehension (249 points, 35%), reading comprehension (249 point, 35%), cloze or error correction (70 points, 10%), and writing and translation (142 points, 20%).

The NCETC has taken a variety of measures to disseminate the newly adopted score reporting system. The CET-4 Syllabus presents a general description of scoring principles, which are further elaborated on its official CET website. A Table comparing the report score and percentile in the CET-4 is posted on the website to facilitate both test takers' and score users' understanding of an examinee's English proficiency and percentile position in any CET-4 a test taker took. In addition, the NCETC explained and interpreted the scoring system on its teaching reform conferences held in different cities to make teachers have a better understanding of it (Jin, 2006).

In discussion of score-based decisions in Chapter 8, questionnaire findings identified different attitudes between students and teachers to the SRF. Interviews helped to reveal the reasons underlying this discrepancy. As discussed earlier, the less satisfaction from students resulted from their lack of knowledge on score equation and transformation of a norm-referenced test. Even though both students and teachers acknowledged that the SRF is more informative and diagnostic than the old certificate, which can be viewed as a successful reform measure, there is room for further improvement in its clarity and understandability. Several student interviewees commented that it would be desirable if descriptions about language use ability corresponding to certain score range were presented. Take IELTS for example, a 9-band scale is created, corresponding to each level of English competence, which is verbally described. Thus, it is advisable that the NCETC provide verbal descriptions or can-do statements to different score range so that students' test performances can be more meaningful.

Timeliness of SRF

The CET-4 is held twice annually, in June and December. The authorized score reporting websites release test takers' score profiles two months after each test administration, usually at the end of August and February. Compared with IELTS which posts test takers' scores 13 calendar days after the test, and TOEFL iBT, releasing test scores online 10 days after the test date, it seems that the NCETC cannot be described as prompt in releasing CET scores. However, given its 10 million test takers every year, this period of time is acceptable. In addition, students usually get to know their scores during the summer and winter vacations, which is still in time for them to make decisions on whether to take the CET-6 and how to adjust their learning activities in a new academic semester based on their performances.

Test takers' official Score Report Forms are delivered to EFL program administrators in April and October, which usually does not cause time insufficiency for most universities to make decisions related to potential test takers in future and long-term teaching curriculums, since the CET-4 is designed to evaluate students' English proficiency after their completion of the foundation stage of study. However, for universities intending to stop normal teaching and open test preparation courses for CET-4 repeaters and CET-6 test takers, they may feel short of time. Take U1 for example, once the Score Report Forms are issued to university, the number of students below and above the cut score of 425 will be calculated. In the fourth semester, usually at the beginning of May, students will be placed into different test preparation courses based on the cut score. EFL program administrators only have one month or so for this placement decision. Just as the administrator in U1 said:

> Time is urgent for us to open test preparation courses and rearrange all the related resources including classrooms, schedules and teachers, because students in one intact class will be classified and then merged with students from other classes. However, we understand that it takes time for the NCETC to issue and deliver the Score Report Forms to universities.
>
> (Administrator of U1, personal communication, June 8, 2010)

One point needs to be clarified. Such time insufficiency should not be attributed to the CET-4, since the decision to open test preparation courses is not what the

CET-4 is intended for. Thus, overall speaking, CET-4 scores are released in a timely manner to stakeholders.

9.4 Warrant on washback

Warrant 3: The consequences of using the CET-4 and of the decisions made on it help promote desirable instructional practices and effective learning in College English instructional settings.

Evidence: This warrant can be supported by the survey results revealed by the NCETC in their validation study (Yang & Weir, 1989), in which both teachers and students showed positive attitudes to beneficial consequences brought about by using the CET-4 (Yang & Weir, 1998). The Ministry of Education also acknowledged the beneficial contributions the CET-4 has made to college English teaching and learning since its launch in 1987 (Wu, 2005; Zhang, 2008). Universities also benefit from the administration of the CET-4, since the NCETC provides score package of students and the relevant statistics of their test takers so that EFL program administrators can make accurate evaluations and appropriate decisions pertaining to English teaching and learning situations in their universities.

Rebuttal: The use of the CET-4 generated unintended or negative consequences (such as the phenomenon of "teaching and learning to the test", narrow of curriculum, anxiety).

In Chapter 4 I have discussed that based on my review from scholarly articles warrant 3 was found to be open to challenges and criticism. Therefore, the above rebuttal was proposed in accordance with the negative washback of the CET-4. In addition, as to the revised version launched at the end of 2006, the NCETC only gave a general statement that stakeholders reacted positively to the trial version of the test and the reform also received favorable media coverage (Jin, 2006). No detailed evidence from the NCETC is available on feedback from both test takers and test users of the revised CET-4. No direct reports are released on how they actually benefit from administration of the test. This may also become the potential rebuttal to the test. Thus, the following part sets out to explore how the major groups of stakeholders view the consequences of using the CET-4. Meanwhile, criticisms against the CET-4 and the corresponding rebuttal were identified and listed above. Findings generated from questionnaires and interviews will determine whether the warrant pertaining to the beneficial washback can be supported or refuted.

9.5 Backing from descriptive statistics of questionnaires and interviews

As noted earlier, 128 teacher questionnaires and 753 student questionnaires were kept as valid for data analyses. There are overlapping themes in both questionnaires to explore any attitudinal differences, so responses from teachers and students on similar items were grouped together in the following report. Findings were described on classified categories rather than on the original item orders in the questionnaires. These categories mainly include general perceptions of the CET-4, evaluations of its design, test preparations, test-taking activities, and the College English teaching and learning practices.

9.5.1 General perceptions of the CET-4

It is important to explore how stakeholders perceive an assessment because their perceptions are believed to have explicit and implicit influences on their teaching and learning qualities (Brown & Hirschfeld, 2008, p.3; Entwistle, 1991). Thus, this section is to investigate stakeholders’ general perceptions of the CET-4, mainly including teachers’ perceptions of and attitudes to the revised CET-4, students’ test-taking motivations, the overall influences of the test on College English teaching as well as specific influences on teachers and students respectively.

Teachers' perceptions of the post-2006 CET-4

Since student respondents recruited in the main study only took the post-2006 CET-4 and might have no idea about pre-2006 test design, several categories of questionnaire items were designed specifically to explore how teachers perceived the major reasons behind the 2006 CET-4 reform, revisions in test paper, and possible corresponding changes in their teaching.

Table 9.1 presents teachers' responses to the major reasons behind the 2006 CET-4 reform according to the descending mean values. The top three major reasons were "To motivate students to lay more emphasis on listening ability", "To further improve the CET-4 as a measure of students' English proficiency" and "To meet social needs for graduates with higher English proficiency". The rank indicated that teachers were clearly aware of the reform reasons and agreed with the intended purposes of this reform. As noted earlier, the significant change of the post-2006 CET-4 is the listening component, which adopts new task types and content in accordance with the CECR to prioritize development of students' listening ability. In addition, with mean values all beyond 3.50, teachers' overall ratings in this category were positive, indicating that they had a good understanding of the theoretical underpinnings and practical concerns behind the test reform.

Table 9.1Teachers' perceptions of major reasons behind the 2006 CET-4 reform

Item	Mean	SD
• To motivate students to lay more emphasis on listening ability	4.14	.897
• To further improve the CET-4 as a measure of students' English proficiency	3.83	.860
• To meet social needs for graduates with higher English proficiency	3.78	.841
• To refine testing methods	3.72	.861
• To positively impact the college English teaching and learning	3.68	.820
• To meet the demands of tertiary education	3.55	869
• To prepare students for their future career	3.50	.956

When further asked about their perceptions of changes in test paper (see Table 9.2). Teachers evaluated that the revised test put a lot more emphasis on listening (Mean=4.14) and integrated skills (Mean=3.95). In other words, they believed the test tasks were designed more communicatively oriented. Inclusion of more constructed response items also reflected more emphasis on examining students' productive skills. Teachers also thought that there was a slight improvement on measuring reading skills and on enhancing authenticity. Since the post-2006 CET-4 cancelled the traditional vocabulary and structure component, teachers agreed grammatical knowledge was less emphasized (Mean=2.50), but the emphasis on vocabulary knowledge remained relatively unchanged (Mean=3.09).

Table 9.2 Teachers' perceptions of changes in the revised CET-4 paper

Item	Mean	SD
• Emphasis on listening	4.14	1.055
• Emphasis on integrated skills	3.95	.775
• Emphasis on being communicatively-oriented	3.66	.890
• Emphasis on productive skills	3.62	.861
• Emphasis on reading	3.50	.860
• Emphasis on authenticity	3.49	.813
• Emphasis on vocabulary knowledge	3.09	1.004
• Emphasis on grammatical usage	2.50	.939

Since the teaching objective is gauged to emphasize development of students' listening ability, it is assumed by the NCETC that increasing the weight of listening component in the test is likely to bring about more emphasis on listening ability in teaching. It is also predictable that teachers' awareness of this emphasis may have direct effects on their teaching, so another category was designed to explore the possibility of teachers' making corresponding changes in their teaching. Table 9.3 indicates that such an assumption is reasonable.

Teachers agreed they would adopt new teaching methods (Mean=3.82), and teach in accordance with revised test formats and contents (Mean=3.80). More specifically, they strongly agreed that they were likely to use a more communicative teaching approach (Mean=3.95) and lay more emphasis on developing students' listening ability (Mean=4.14). Fast reading, as a new task type, would receive more attention as well (Mean=3.86).

Table 9.3 Corresponding changes likely to be made in teachers' teaching

Item	Mean	SD
• To teach in accordance with the new test formats & contents	3.80	.920
• To adopt new teaching methods	3.82	.757
• To use a more communicative teaching approach	3.95	.835
• To lay more emphasis on developing students' listening ability	4.14	.885
• To lay more emphasis on developing students' fast reading ability	3.86	.867
• To lay more emphasis on developing students' careful reading ability	3.62	.814

To sum up, findings from the above three categories revealed that teachers were generally supportive of the test reform. They had a clear awareness of the background and emphasis of the test reform. They knew the revised test was intended to direct teaching focus on developing students' communicative competence, particularly listening ability. In addition, teachers' understandings of the revised content, and their implicit agreement with the rationales and intended purposes of the reform increased chances for them to make corresponding changes in their teaching.

Students' motivations for taking the CET-4

When asked about what motivated students to take the CET-4 for, students and teachers produced the following mean list. Table 9.4 presents the results and all the items ranked in a descending order of students' mean values.

Both groups listed "To get the Bachelor's degree" and "To obtain advantage in employment" as the top two. On the one hand, the CET-4 is designed to examine the effect of college English teaching, to examine whether students have reached the curriculum requirements and cultivated the integrative English ability to meet the social demands in future work and social interactions. However, on the other hand, the questionnaire revealed that students were motivated to take the CET-4 mainly for its uses in obtaining academic qualification and career development. Such a belief indicates that the unintended purposes to some extent have distorted the intended purposes of the CET-4.

Table 9.4 Students' motivations for taking the CET-4

Item	Group	Mean	SD
To get the Bachelor's degree	Students	4.21	.815
	Teachers	4.07	.924
To obtain advantage in employment	Students	3.82	1.149
	Teachers	4.20	.754
To satisfy academic credit requirement	Students	3.49	1.150
	Teachers	3.90	.884
To check my English proficiency	Students	3.22	1.178
	Teachers	3.14	.945

Test consequences

When teachers were asked to give responses to the overall consequences of the CET-4 in the past two decades, they listed two items related to the negative washback of the CET-4 at the top (see Table 9.5). Teachers moderately agreed that the CET-4 in the past decades had induced the phenomenon of "teaching and learning to the test" (Mean=3.55), and "higher marks, lower abilities" (Mean=3.42). Meanwhile, they also agreed that the administration of the test had positive influences in "Promoting college English teaching" (Mean= 3.31) and "Improving students' linguistic proficiency" (Mean=3.30). However, they expressed uncertain attitude on whether the test helped improve students' communicative competence (Mean=3.14). It can be inferred that in teachers' opinions the negative consequences of the CET-4 outweighed its positive.

Table 9.5 Teachers' perceptions of the CET-4 influences in the past decades

Items	Mean	SD
• Inducing the phenomenon of "teaching and learning to the test"	3.55	.962
• Inducing the phenomenon of "high marks, low abilities"	3.42	1.024
• Promoting college English teaching on the whole	3.31	.945
• Improving students' linguistic competence	3.30	.807
• Improving students' communicative competence	3.14	1.010

The next two categories specifically explored how CET-4 scores influenced students and teachers affectively (see Tables 9.6 and 9.7). From Table 9.6 we can see that students thought their CET-4 scores to some extent had influences on their self-confidence, self-evaluation, sense of achievement, and study interest, but they were uncertain about its influences on their images among classmates and teachers. In their interviews, students explained that once they passed the cut score of 425, they viewed it as a success in their English study and felt more confident in their English proficiency. In addition, they would not judge their classmates only based on their CET-4 scores since the overwhelming majority of students could pass the test upon graduation.

Table 9.6 Aspects where CET-4 scores influence students

Item	Mean	SD	Item	Mean	SD
Self-confidence	3.58	1.207	Study interest	3.42	1.320
Self-evaluation	3.51	1.095	Image among my teachers	2.86	1.219
Sense of achievement	3.43	1.228	Image among my classmates	2.78	1.213

As Table 9.7 shows, teachers admitted that students' CET-4 scores, usually the passing rates of the whole class to some extent influenced their sense of achievement and self-evaluation. They were uncertain about influences of CET-4 passing rates on their popularity with students and image among colleagues. They thought the passing rate only had a slight influence on their academic promotion and cash bonus.

Table 9.7 Aspects where CET-4 scores influence teachers

Item	Mean	SD	Item	Mean	SD
Sense of achievement	3.54	1.108	Image among colleagues	3.05	1.113
Self-evaluation	3.45	1.114	Academic promotion	2.69	1.278
Popularity with students	3.09	1.080	Cash bonus	2.17	1.198

According to teachers' interviews, the four sampled universities all had the policy that the CET-4 passing rate would be referred to in teachers' academic promotion. However, published journal articles actually serve as a decisive factor. As to cash bonus, teachers in U1 said their university once gave cash award to teachers whose CET-4 passing rate was beyond 80%. With the increasing passing rate in recent years, this awarding policy had been ceased.

9.5.2 Evaluations of the CET-4

Students' conceptions of assessment are believed to have a significant impact on the quality of learning ((Brown & Hirschfeld, 2008, p.3; Entwistle, 1991). Therefore, this section investigates how students and teachers think of the CET-4 design, including their perceptions of test difficulty, the extent to which it serves as an indicator of students' English proficiency, and in particular the design of listening and reading components. Before discussing the above aspects, I will report how students viewed the importance of the basic language skills and how they evaluated their own English proficiency in order to provide a benchmark to weigh their evaluations of the test design.

Table 9.8 presents the descriptive data on students' responses to importance of the five skills (in an order of importance): listening, speaking, reading, writing and translation. It can be seen that the overwhelming majority of students thought that listening (84.9%) and speaking skills (74.8%) were the most important. Reading skill, once emphasized as the most essential, only ranked as moderately important (44.2%). Writing (65.8%) and translation skills (82%) were labeled as the least important skills to cultivate. The promising indication is that students were aware of the significance of productive skills.

Table 9.8 Students' perceptions of the importance of language skills

Skills	The most important		→	The least important	
	1	2	3	4	5
Listening	43.1%	41.8%	9.8%	4.4%	1.8%
Speaking	37.8%	37.0%	10.5%	7.3%	7.3%
Reading	14.3%	10.2%	44.2%	20.8%	10.5%
Writing	3.0%	7.3%	23.9%	44.4%	21.4%
Translation	1.8%	3.7%	12.5%	23.0%	59.0%

However, when students were asked to evaluate their own English proficiency in the above five domains, the list changed greatly (in an order of strength): reading, writing, translation, listening and speaking (see Table 9.9). A vast majority of students (about 69.8%) evaluated their reading ability as the strongest or stronger. About half of the respondents ranked their listening (46.2%) and translation skills (52.8%) as moderate or poorer. About 41% of students viewed speaking as their poorest skill.

Table 9.9 Students' self-assessed English proficiency

Skills	Strongest	→		Poorest	
	1	2	3	4	5
Listening	14.1%	16.0%	16.4%	29.8%	23.7%
Speaking	9.2%	11.6%	16.4%	21.8%	41%
Reading	46.8%	23.0%	17.9%	8.2%	4.1%
Writing	23.4%	32.7%	18.9%	17.8%	7.3%
Translation	6.4%	16.8%	30.5%	22.3%	23.9%

The sharp contrast in the above two tables indicated that on the one hand, students believed that listening and speaking abilities should be the priorities to cultivate so that they could communicate effectively in English. On the other hand, students were aware of their weaknesses in both. It can be inferred that there is a divergence between what students need to cultivate and what is actually taught and emphasized in College English teaching and learning. In this sense, it is a positive and appropriate reform measure that the current CET-4 increases the weight of listening and reduces that of reading with the purpose to exert more positive washback on College English teaching. It also evidences that the CET-4 reform is necessary and prompt conforming to both students' needs and social needs.

Stakeholders' evaluations of test difficulty

Students and teachers were asked to evaluate the difficulty of the test tasks on a 5-point Likert scale from "very easy" to "very difficult" (see Table 9.10).

Table 9.10 Stakeholders' evaluations of test difficulty

Test task	Group	Mean	SD
The overall CET-4	Students	3.80	.645
	Teachers	3.30	.538
Listening component	Students	3.49	.766
	Teachers	3.61	.643
Short conversations	Students	2.79	.749
	Teachers	2.73	.704
Long conversations	Students	3.43	.768
	Teachers	3.38	.689
Listening passages	Students	3.75	.751
	Teachers	3.65	.683
Compound dictation	Students	3.79	.816
	Teachers	3.66	.806
Reading component	Students	3.61	.698
	Teachers	3.21	.572
Fast reading	Students	3.67	.871
	Teachers	3.17	.795
Banked cloze	Students	3.74	.832
	Teachers	3.45	.802
Careful reading	Students	3.51	.706
	Teachers	3.35	.705
Cloze	Students	3.51	.753
	Teachers	3.31	.624
Translation	Students	2.94	.715
	Teachers	3.12	.687
Essay writing	Students	2.89	.645
	Teachers	3.19	.673

In order to have a direct visual picture, two bar charts were produced to facilitate the interpretations. First, evaluations of the overall CET-4 and its main components were examined. According to Figure 9.1, students' rating on the overall test almost reached difficult level (Mean=3.80), while teachers' rating was just above the average difficulty (Mean=3.30). In terms of each component, teachers listed listening as the top difficult (Mean=3.61), cloze (Mean=3.31) as the secondary, followed by reading component (Mean=3.21). In contrast, students ranked reading as the top difficult (Mean=3.61), cloze (Mean=3.51) and listening component (Mean=3.49) as the secondary. Both groups rated writing and translation as comparatively easy ones.

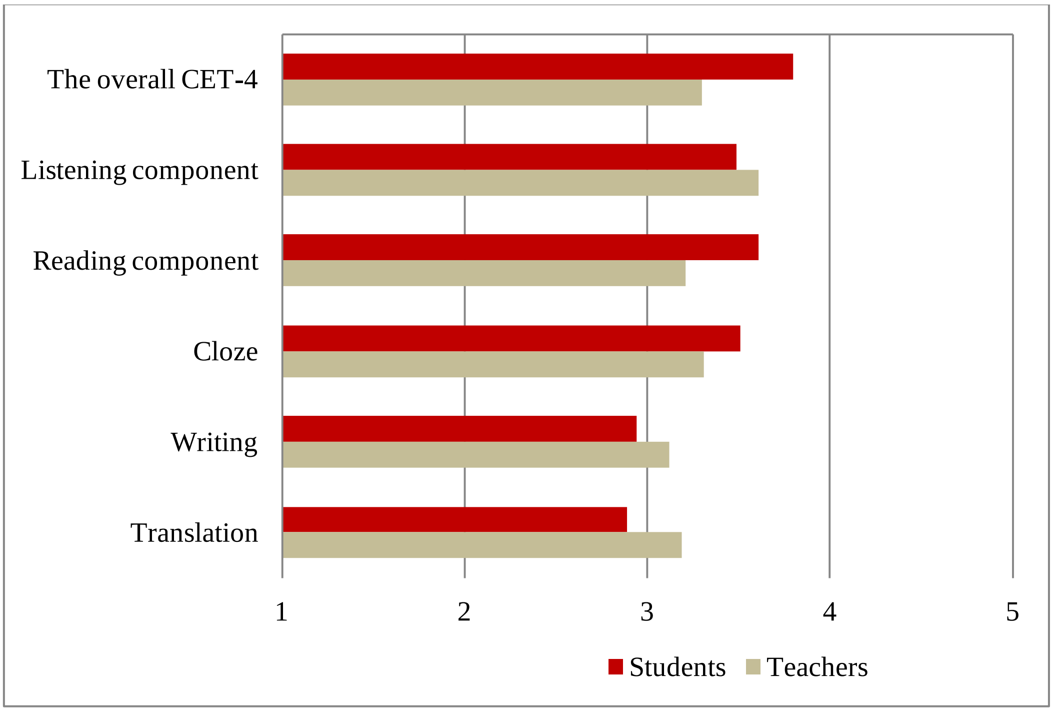

Figure 9.1 Test difficulty of each component

Figure 9.2 displays rating scales on difficulty level of each task type in listening and reading components. A general look at the figure reveals that students and teachers had similar ratings. For example, they all rated compound dictation as the most difficult and short conversations as the easiest among all the task types. Listening passages, banked cloze and careful reading passages were rated as almost equally difficult. The major contrast was observed in their evaluations on fast reading section (Teacher mean=3.17, Student mean=3.67). Students thought that fast reading was no less difficult than listening passages and banked cloze.

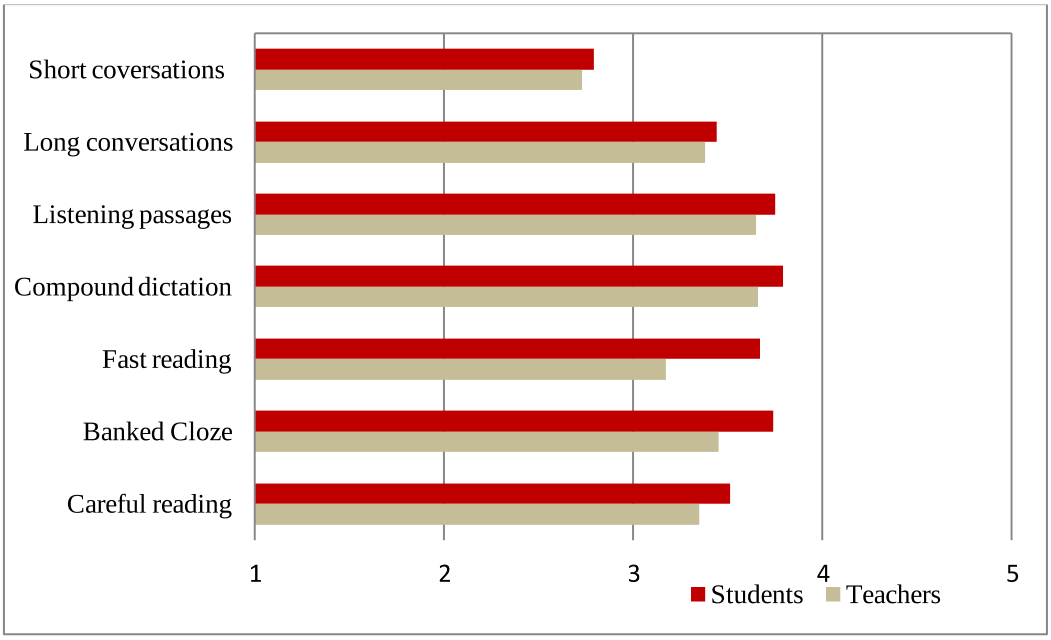

Figure 9.2 Test difficulties of task types in listening and reading components

The follow-up interviews helped get further explanations. The compound dictation was rated as the most difficult because students were required to produce output. Student interviewees expressed that they did not have much difficulty writing down the first seven dictated words. However, for the last three blanks, even though they could understand the sentences, they failed to jot down them completely and correctly within the short pause. As to short conversations, students explained that they could understand the dialogues between two speakers without using test-taking strategies. In addition, having done plenty of exercises, they were familiar with the stereotyped themes and did not have much difficulty in knowing what the item was intended to test. For example, there would be questions asking, “Where did this conversation take place?” and “What is the relationship between the two speakers?” They could get clues by catching words like menu, book, and assignment to figure out the correct choices. Therefore, they rated short conversations as the easiest. With regard to reading tasks, careful reading was viewed as more challenging, requiring large vocabulary and the ability to handle longer sentences. In terms of fast reading, both teachers and students agreed that the passage was not so difficult to understand, but students thought that time pressure increased the difficulty in getting better performances in this task.

Stakeholders' evaluations of the test design

Students and teachers were also asked to evaluate "To what extent the CET-4 can serve as an accurate indicator of students' overall English proficiency". Table 9.11 presents the mean list of their evaluations. This table is quite revealing in several aspects. First, teachers gave higher ratings almost to all the task types than students did, indicating their more favorable evaluations of the CET-4 qualities. Second, in terms of the main test components, both teachers and students listed writing, translation and listening components as the top three tasks that could better measure students' overall English proficiency. Third, with regard to the task types in listening and reading components, teachers believed long conversations (Mean=3.81) and compound dictation (Mean=3.81) could better reflect test takers' overall English proficiency. Students gave almost the same ratings to three listening tasks. Students' lower ratings on reading tasks revealed their uncertainty about whether fast reading (Mean=3.12) and banked cloze (3.17) could reflect their proficiency. The above findings indicate that tasks assessing productive skills like listening, translation and writing components, or constructed response items requiring language output were evaluated more favorably than the MCQ. These tasks types, being more communicatively oriented and demanding for integrative abilities, thus have higher ratings as accurate indicators of test takers' proficiency.

To summarize Tables 9.10 and 9.11, we can see that teachers and students regarded reading and listening components as comparatively more difficult tasks, but listening component was viewed as more communicatively oriented. Writing and translation were regarded as the easier tasks but still could be more effective to elicit students' performances of English communicative competence.

Table 9.11 Evaluations of the CET-4 as an accurate indicator of students' overall English proficiency

Test task	Group	Mean	SD
The overall CET-4	Students	3.22	1.136
	Teachers	3.55	.840
Listening component	Students	3.54	1.038
	Teachers	3.84	.778
Short conversations	Students	3.54	1.004
	Teachers	3.62	.774
Long conversations	Students	3.52	.972
	Teachers	3.81	.718
Listening passages	Students	3.51	1.011
	Teachers	3.70	.847
Compound dictation	Students	3.55	1.075
	Teachers	3.81	.876
Reading component	Students	3.43	1.070
	Teachers	3.73	.778
Fast reading	Students	3.12	1.185
	Teachers	3.68	.896
Banked cloze	Students	3.17	1.124
	Teachers	3.61	.825
Careful reading	Students	3.38	1.045
	Teachers	3.68	.752
Cloze	Students	3.12	1.136
	Teachers	3.59	.883
Translation	Students	3.63	1.011
	Teachers	3.77	.889
Essay writing	Students	3.97	.932
	Teachers	3.87	.942

Teachers and students were further asked about test design of listening and reading components. Table 9.12 ranks the related items according to students' descending mean values.

Table 9.12 Evaluations of test design of listening and reading components

Item	SM	TM
• Printing listening questions on the paper can facilitate students' understandings.	4.24	3.62
• Adding titles to passages can facilitate students' prediction of the content.	3.94	3.78
• Reading passages cover a wide range of topics such as humanities, culture, history, education, geography, and science, etc.	3.70	3.84
• Reading genres should be more diverse with practical passages such as letters, advertisements, and instructions besides the dominant genres of argumentation, narration and exposition.	3.41	3.88
• Listening tasks are similar to those in the real life situation.	3.01	3.21
• Reading tasks are similar to those in the real life situation.	2.97	3.23
• The overall formats and content of the CET-4 are satisfactory.	2.69	3.27

Notes. SM= student mean; TM= teacher mean.

The top two items indicate that students expected to read questions before their listening in order to listen with prompts, or predict reading content by titles of passages. In other words, we may assume that predicating with certain clues can greatly facilitate test takers' understanding of listening and reading input. Students and teachers showed higher level of agreement with topic distributions of reading passages but suggested that reading genres should be more diverse. In terms of the authenticity of test tasks, both groups showed their uncertainty. Since authenticity plays an important role in shaping stakeholders' positive perceptions of a test, the lower ratings on this test quality deserves test developers' attention. Students held particularly lower level of agreement with the overall design of the CET-4 (Mean=2.69) than teachers did (Mean=3.27). It may indicate that the test content and format of the CET-4 still have room for further improvement. The lower rating may also be attributed to students' mixed feelings to the test, or specifically, dissatisfaction with pressure and anxiety caused by the test.

9.5.3 Test preparation for the CET-4

This section describes stakeholders' test preparation activities. When asked about which semester they started to prepare for the CET-4, 59.5% of the students responded their test preparations started from the third semester in which they were going to take the test. Over 20% started earlier at the second semester. About 30% of teacher respondents reported that test preparation in their universities started from the second semester, about 49% reported from the third semester.

In terms of the effect or the outcome of test preparations, students and teachers shared similar ideas. As shown in Table 9.13, both of them agreed that test preparations to some extent could improve students' CET-4 scores, but such preparations were less effective in improving students' overall English proficiency.

Table 9.13 Effect of test preparations

Item	Group	Mean	SD
Test preparations can improve students' scores.	Students	3.87	1.056
	Teachers	3.81	.984
Test preparations can improve students' overall English proficiency.	Students	3.41	1.163
	Teachers	3.43	1.026

When students were asked to rank their test preparations methods, the list was produced in an order of importance (see Table 9.14): to memorize vocabulary, to learn test-taking strategies, to do exercises, to review textbook, to attend coaching schools.

In their interviews, students explained that they spent most of time memorizing words. For one thing, vocabulary served as a foundation for their understanding of the test content. For another, they found that they might not improve their listening ability but could enlarge their vocabulary within a short period of time. "To learn test-taking strategies" was put at the secondary place because students felt that some strategies were effective in helping them eliminate distractors. None of the interviewees reported they attended coaching schools since their universities would stop normal teaching for test preparation courses. Above all, students thought that their self-preparations for the test were more effective.

Table 9.14 Students' evaluations of test preparation methods

Test preparation methods	The most important → The least important				
	1	2	3	4	5
To memorize vocabulary	45.0%	27.4%	18.1%	6.6%	2.8%
To learn test-taking strategies	26.5%	29.7%	27.3%	12.4%	4.1%
To do exercises	18.1%	23.2%	23.2%	23.8%	11.6%
To review textbook	6.5%	12.4%	18.5%	28.9%	33.6%
To attend coaching schools	4.1%	7.2%	12.6%	28.4%	47.8%

As discussed earlier, to open test preparation course is a common practice in the university community of Xi'an. Students and teachers were asked about their opinions on relationship between teaching and test preparations. Table 9.15 illustrates their responses to test preparation course.

Table 9.15 Attitudes to test preparation courses

Item	Group	Mean	SD
• It's necessary to stop the normal teaching for test preparation.	Students	3.20	1.135
	Teachers	2.92	.985
• I learn/teach to the test during the semester in which I/my students take CET-4.	Students	3.19	1.144
	Teachers	3.32	1.064

Compared to students who agreed to replace the normal teaching with test preparation (Mean=3.20), teachers showed uncertain attitudes (Mean=2.92). However, both of them agreed that they learned or taught to the test at the closer time of the CET-4 administration.

Table 9.16 displays the results of students' and teachers' evaluations of test preparation activities in class on a 5-point Likert scale of frequency (1= never, 5=

always). Except the last item, the mean values given by students were around 3.3 while those from teachers were around 3.6. It seems that test preparation class involved all the activities. Apparently, the higher values from teachers indicate that they believed they conducted these activities more frequently in class. The sharp contrast was found in the last item, which had lowest value from students but highest from teachers. Teachers maintained that their teaching was still committed to improving students' five language skills even though they delivered test preparation classes. However, in students' eyes, the most frequent activity was to listen to their teachers' analyzing and explaining authentic or mock test papers.

Table 9.16 Evaluations of the frequency of test preparation activities

Item	Group	Mean	SD
• Explaining authentic or mock test papers	Students	3.38	1.024
	Teachers	3.66	.907
• Developing students' test taking strategies	Students	3.31	.999
	Teachers	3.62	.824
• Offering information about test contents and the formats	Students	3.28	.995
	Teachers	3.66	.863
• Doing exercises related to the five language skills	Students	**2.78**	1.106
	Teachers	**3.83**	.795

When asked about the medium of instruction teachers used in test preparation courses, as Figure 9.3 shows, about half of students and one third of teachers chose "English with occasional Chinese", while 40% of them reported to use "half English and half Chinese". It is understandable that using Chinese can facilitate teachers' expressions on and students' accurate understandings of the test design, task characteristics, application of test-taking strategies and so on, but English input students received during test preparation classes definitely decreased.

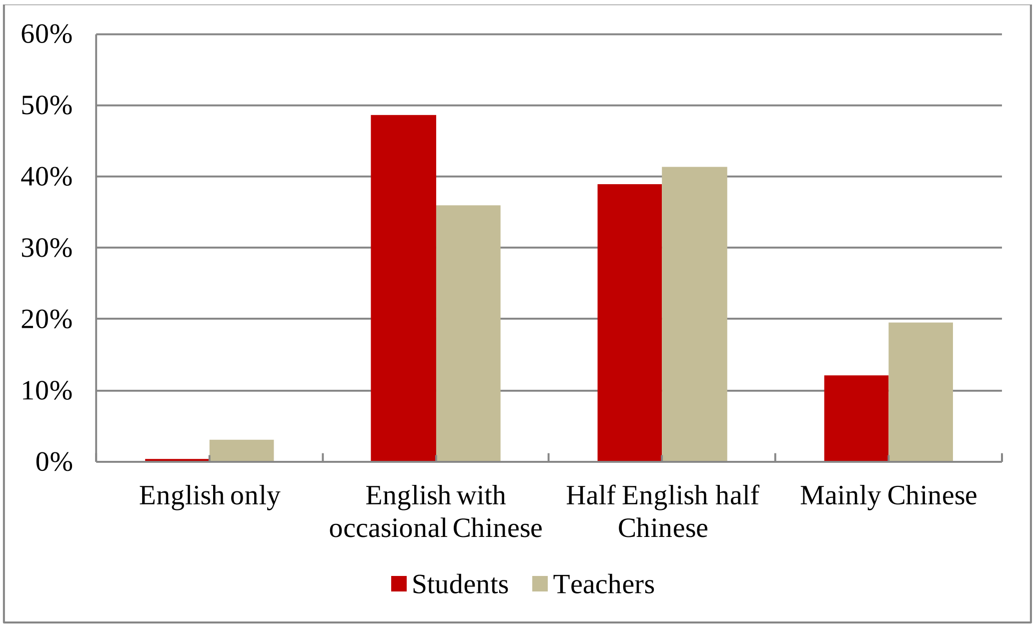

Figure 9.3 Medium of instruction that teachers use in test preparation course

After getting a basic picture of classroom activities in test preparation course, I intended to investigate students' test preparation outside of the class. Students were asked to evaluate the average amount of time they spent on different test preparation activities per week. Teachers were also asked to make evaluations based on their understanding of students' learning behaviors. Table 9.17 presents the results.

Table 9.17 An average amount of time spent on test preparation activities per week

Item	G	M	SD	Item	G	M	SD
Doing authentic test papers	S	3.55	1.035	Practicing writing	S	2.69	1.030
	T	3.26	1.101		T	3.09	.896
Practicing listening	S	3.42	1.142	Practicing translation	S	2.58	1.032
	T	3.45	.954		T	2.97	.887
Memorizing vocabulary	S	3.37	1.137	Learning test taking strategies	S	2.49	1.030
	T	4.07	.821		T	3.23	.997
Practicing reading	S	3.30	1.069	Practicing speaking	S	1.86	1.016
	T	3.59	.798		T	2.41	1.118
Doing mock tests	S	3.11	1.181	Reviewing grammar	S	1.8	.856
	T	3.56	.942		T	2.60	.940

Note. G=group; M= mean; S=students; T=teachers.

All the items ranked according to students' descending mean values. The left column contains the top five items whose mean values given by students were about 3.5, indicating students spent an average of three to four hours per week on each of these activities. "Doing authentic test papers" (Mean=3.55) was listed at the top, followed by "Practicing listening" (Mean=3.42). Since the CET-4 increased the weight of listening component, the immediate result is the time students allocated on practicing listening was more than on reading. On the other hand, teachers' rankings of the five items were very different from those of students. For instance, teachers assumed that students would spend most of time memorizing vocabulary (Mean=4.07), but students only listed it at the third place (Mean=3.37). Students listed "Doing authentic test papers" (Mean=3.55) at the top, while teachers believed that students would spend more time doing mock test papers (Mean= 3.56) than authentic ones (Mean= 3.26). The possible reason might be teachers were asked to evaluate students' test preparation activities outside of class based on their assumptions, so discrepancies were avoidable. If we further examine the right column, we can see that students spent about one hour per week in practicing productive skills of writing and translation, which only occupied 20% of weight in the CET-4. Since the speaking test is not compulsory to students, and the task type directly examining grammar is cancelled, we can see that students did not bother to spend time on them. Findings from this category revealed a phenomenon that students tended to spend more time on skills examined in the test and ignored those untested.

It is interesting to notice that students rated memorizing vocabulary as the most important test preparation method, but spent more time doing authentic test papers, so in their interviews this question was further brought up. They explained that memorizing vocabulary was effective but it served as the dominant method at the earlier stage of test preparation. With the test approaching, they would spend more time doing authentic test papers to familiarize themselves with the test format. When teachers were asked why they assumed students would spend more time on mock test papers rather than authentic ones, they explained that since the revised

CET-4 was launched at the December 2006, up to December 2009 only six authentic test papers were available. Therefore, they assumed students would buy mock test papers. Students offered their explanations of higher ranking of authentic test paper. In spite of different formats and content, they did the authentic papers of the pre-2006 CET-4 because they believed that the old versions were still useful in estimating the test difficulty and diagnosing their own weaknesses.

In the test preparation section, several items were designed related to reviewing authentic or mock test papers. Figure 9.4 displays the number of authentic and mock test papers reviewed by students and teachers respectively. We can see that about 50% of teachers and 35% of students reported they reviewed about five to six sets of test papers, while the same number of teachers (24.2%) and students (25.6%) reviewed 11 to 15 sets.

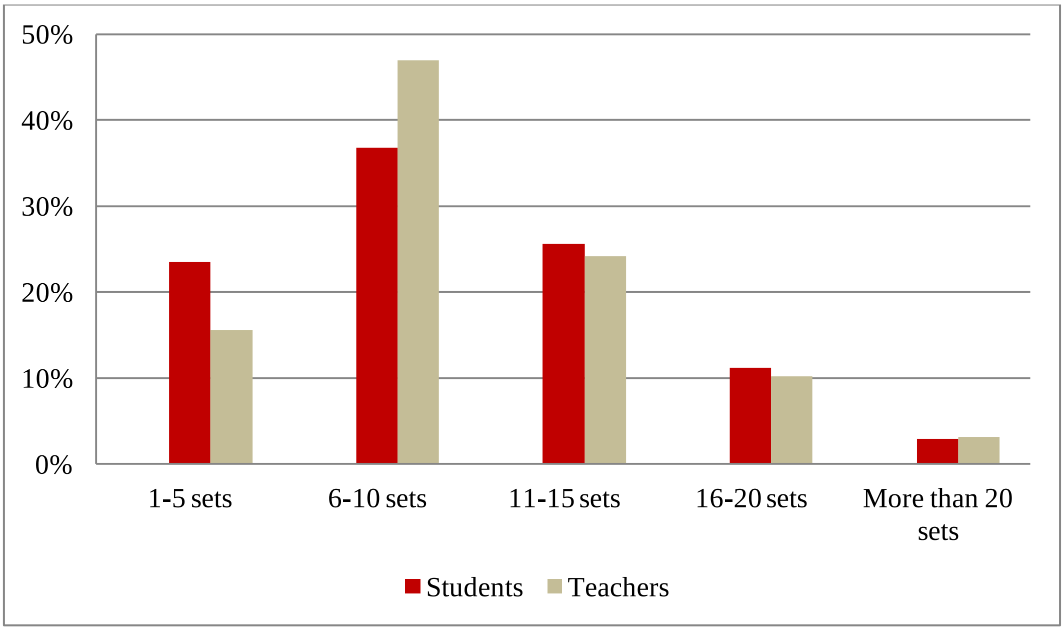

Figure 9.4 The number of test papers reviewed for CET-4 preparations

Since reviewing test papers seemed to occupy most of time, teachers and students were further asked about the functions of doing test papers. From Table 9.18, we can see that both groups strongly agreed that doing test papers could familiarize students with the CET-4 content and format. They also moderately agreed that this method could help students diagnose their learning strengths and weaknesses.

Table 9.18 Functions of doing test papers

Item	Group	Mean	SD
• Doing past or mock test papers can familiarize me/students with the CET-4 content and format.	Students Teachers	4.15 4.14	.782 .801
• Doing past or mock test papers helps me/students diagnose their learning strengths and weaknesses.	Students Teachers	3.89 3.77	.864 .883

9.5.4 Test-taking activities

This section is to investigate students' test taking activities, mainly including the possible problems and difficulties affecting their listening and reading performances, and the strategies they adopted to handle them.

Factors that affect students' test performances

Students were asked to evaluate factors affecting their listening and reading performances respectively on a 5-point Likert scale of agreement, and teachers were also asked to make assumptions on possible problems and difficulties their students may encounter. Tables 9.19 and 9.20 present the results according to students' descending mean values.

Table 9.19 displays that both groups listed sentence structures and limited vocabulary as the top two factors. Apparently, teachers showed higher level of agreement with the two items than students did. The indication is that linguistic knowledge constituted the major factor affecting students' listening comprehension. It is interesting to notice that students put "tricky options" at the third place while teachers gave this item the lowest rating. Instead, teachers thought "lack of listening skills or strategies was another obstacle in students' listening comprehension.

Table 9.19 Factors affecting students' listening performances

Item	Group	Mean	SD
• Long and complex sentences	Students	3.88	.783
	Teachers	4.02	.813
• Limited vocabulary	Students	3.64	.948
	Teachers	4.03	.904
• Confusion by the tricky options	Students	3.56	.910
	Teachers	3.65	.839
• Lack of background knowledge	Students	3.53	.894
	Teachers	3.84	.954
• Lack of listening skills and strategies	Students	3.44	.918
	Teachers	3.80	.774

Table 9.20 presents mean values of factors affecting students' reading performances. Students listed "Confusion by the tricky options in spite of my understanding" as the top factor affecting their reading performances, while teachers ranked it as the fifth among the six factors. In their interviews, students explained that despite their understanding of the passage, they sometimes spent quite much time distinguishing the confusing and tricky distractors. As a result, they did not have sufficient time to finish all the test items. Teachers rated limited vocabulary as the top one factor, followed by "Long and complex sentence" and "Slow reading speed".

Table 9.20 Factors affecting students' reading performances

Item	Group	Mean	SD
• Confusion by the tricky options in spite of my understanding	Students	4.03	.816
	Teachers	3.82	.873
• Limited vocabulary	Students	3.74	.952
	Teachers	4.02	.837
• Slow reading speed	Students	3.60	1.008
	Teachers	3.90	.744
• Long and complex sentences	Students	3.55	.962
	Teachers	4.01	.779
• Lack of reading skills and strategies	Students	3.40	.960
	Teachers	3.84	.729
• Lack of background knowledge	Students	3.30	.937
	Teachers	3.69	.929

When students and teachers were asked to evaluate the extent to which factors affecting students’ overall CET-4 performances (see Table 9.21), they ranked “Time pressure”, “Difficulty of questions”, and “Difficulty of language” as the top three factors.

Table 9.21 Factors affecting students’ overall CET-4 test performances

Item	Group	Mean	SD
Time pressure	Students	4.47	.921
	Teachers	3.78	.955
Difficulty of questions	Students	3.92	.939
	Teachers	3.71	.962
Difficulty of language	Students	3.73	1.019
	Teachers	**3.90**	.828
Lack of test-taking strategies	Students	3.45	1.066
	Teachers	3.51	.947
Unfamiliarity with topics	Students	3.37	1.136
	Teachers	3.63	.901
Test anxiety	Students	3.06	1.325
	Teachers	3.60	.925

Almost all the student interviewees expressed that they were short of time in completing all the items, particularly in doing fast reading and careful reading parts. Quite a number of students said they left bank cloze at the last place and would blindly pick out options if time were insufficient. They tended to prioritize careful reading passages due to its larger weight. The top two factors listed by students are relevant to test design, which deserves test developers’ attention. Teachers ranked “Difficulty of language” at the top, indicating their belief that students’ performances mainly depend on their own English proficiency.

To sum up the above three tables, sentence structure, limited vocabulary, tricky distractors, reading speed are the major factors affecting students’ listening and reading performances. Since listening and reading components occupy 70% of the total weight, these factors along with time pressure also have influences on students' overall test performances. Both groups listed "Lack of listening and reading strategies" and "Lack of background knowledge" at the bottom rank. The findings seem to indicate that students believed that they were not lack of test-taking strategies. In addition, the selected passages did not favor a particular group of test takers. However, whether these factors would affect test takers' performances as perceived would be discussed in the latter part of this chapter. Inferential statistical procedures were applied to explore to what extent the above factors would influence their test performances.

Students' test-taking strategies

In the above section, I have explored major difficulties students are confronted with in test-taking process. This section focuses on how students cope with the problems. I will discuss teachers' teaching of these strategies along with students' application of them.

Table 9.22 presents students' test-taking strategies in taking the listening component. "Reading options first" and "skipping unknown words" received students' strongest agreement. It indicates that predicating listening materials and keeping full concentration was believed to be important strategies to facilitate students' listening comprehension. Students moderately agreed that they would pay attention to conjunctions (M=3.46) and were uncertain about whether they could catch the interlocutor's stress and intonation (M=3.01) to make judgments. Students were found to seldom take notes to help their memory (M=2.70).

Table 9.22 Test-taking strategies in doing listening component

Item	Mean	SD
• I read options first to predict what I am going to hear.	3.76	1.010
• I just skip unknown words so as to concentrate on the whole.	3.72	.973
• I pay attention to conjunctions such as "but, so that…" to infer speakers' opinions.	3.46	1.014
• I pay attention to the speakers' stress and intonation to infer their intentions or attitudes.	3.01	1.140
• I take notes to help my memory.	2.70	1.153

Correspondingly, a category was designed in teacher questionnaire exploring whether they encouraged and taught students to learn and use these test-taking strategies to improve their listening performances. As shown in Table 9.23, all the items received strong agreement with their mean values around 3.9. It indicates that teachers emphasized all the strategies. A closer look at Tables 9.22 and 9.23 reveals a contrast. "Taking notes to help remember details" received strongest agreement from teachers but least agreement from students. Likewise, "Predicting listening content by reading options first" received teachers' least agreement but students' strongest agreement. In their interviews, teachers explained that they viewed both as necessary, but taking notes while listening is a skill to be encouraged more. On the contrary, students thought listening required high concentration. They tried this strategy, but found taking notes would distract them from what they were listening. Moreover, they failed to jot down as many words as possible. Reading options first was more effective for them to predict the scenario to be heard.

Table 9.23 Teaching of test-taking strategies in doing listening component

Item	Mean	SD
• Taking notes to help them remember details	4.08	.866
• Skip the unknown words so as to concentrate on the whole	4.00	.784
• Paying attention to some conjunctions to infer speakers' opinion	3.95	.767
• Paying attention to stress and intonation to infer speaker's attitudes	3.91	.827
• Predicting the listening content by looking through the options	3.89	.941

Table 9.24 presents students' test-taking strategies in doing the reading component. We can see that scanning was put at the top place, closely followed by "reading questions first". When students came across unknown words, they tended to skip them or guess their meanings from context. They were uncertain whether they would analyze the grammatical structure when encountering complex and difficult sentences.

Table 9.24 Test-taking strategies in doing reading component

Item	Mean	SD
• I scan to search for the specific details.	3.74	.871
• I read questions first before reading passages.	3.71	1.125
• I just skip unknown words and continue to focus on my reading.	3.63	.830
• I guess the meaning of an unknown word in context.	3.48	.881
• I skim to identify the main idea.	3.39	.902
• I guess the meaning of unknown words by its root, prefix or suffix.	2.94	1.037
• I analyze grammatical structures to help me understand complex sentences.	2.86	.993
• I look through the passage first for the main idea before my careful reading.	2.75	1.135

Table 9.25 presents teachers' responses to these strategies. In their teaching, they strongly encouraged using "Skimming and scanning skills for different purposes", and "Guessing unknown words in the context or by word-building knowledge". The major difference between both groups was that students still viewed "Reading questions first" (Mean=3.71) as more effective than "Looking through the passage for the main idea before careful reading" (Mean=2.75).

Table 9.25 Teaching of test-taking strategies in doing reading component

Item	Mean	SD
• Using skimming and scanning skills for different purposes	4.22	.742
• Guessing unknown words in the context or by word-building knowledge	3.96	.827
• Looking through the passage for the main idea before careful reading.	3.84	.858
• Reading questions first before reading passages.	3.74	.982
• Analyzing the grammatical structure for difficult and complex sentences.	3.20	.956

They explained in interviews that they were under great time pressure so that it was impossible for them to look through passages for the first time and then read them carefully. They preferred to read questions first because they knew where to put their focus. Teachers said that they focused more on correct use of reading or listening strategies than on teaching of test-taking strategies for higher scores.

9.5.5 Perceptions of College English teaching and learning

Testing and teaching are closely related (Heaton, 1988). In previous sections stakeholders' perceptions of the CET-4, evaluations of the test qualities, test preparation, and test-taking activities have been discussed. This section delineates a general picture of the status quo of College English teaching and learning. The aspects under investigation include teaching objectives, factors that affect teaching, classroom activities, students' learning motivations, and potential improvement in College English teaching.

When asked about the objectives of College English teaching, an overwhelming majority of teachers agreed (44.5%) and strongly agreed (29.7%) that "The long-term objective is to develop students' ability to use English in an all-round way". Meanwhile, half of teachers also agreed (44.5%) and strongly agreed (13.3%) that "The short-term objective is to help students obtain high scores in the CET-4". The results indicate that the CET-4 has been an important concern in teachers' understanding of the teaching objective due to its high-stakes. It naturally generated a question whether their understandings of the teaching objective could implicitly influence their teaching. The next category thus asked teachers to evaluate influences of these factors on their teaching (see Table 9.26).

Table 9.26 Factors influencing teachers' teaching

Item	Mean	SD	Item	Mean	SD
Teaching experience	3.84	.774	Students' expectations	3.68	.841
Teaching belief	3.81	.936	The CECR	3.61	.737
Past experience as a language learner	3.77	.806	The CET-4	3.53	.944
University's curriculum requirement	3.70	.807	Textbooks	3.42	1.032

The mean values of most items were above 3.50, indicating that teachers moderately agreed that these factors all had various degrees of influences on their teaching. Teaching experience and teaching belief were listed at the top, followed by teachers' past experience as a language learner. The university's curriculum requirement had higher value than the national CECR. This finding evidenced what was discussed in Chapter 8, indicating that institutional decisions on curriculum arrangements tended to have more direct and powerful consequences on teaching and learning activities. The CET-4 (Mean=3.53) was evaluated as the less influencing factor. However, it should be noted that the previous finding has revealed that washback of the CET-4 on English teaching and learning is more obvious and intense at the closer time of the test administration.

Both students and teachers were asked to evaluate the frequencies of the following classroom activities (see Table 9.27).

Table 9.27 Frequency of classroom activities

Item	Group	Mean	SD
• Explaining the textual meaning	Students	4.15	.673
	Teachers	3.91	.743
• Explaining textbook exercises	Students	4.00	.787
	Teachers	3.59	.865
• Explaining language points such as vocabulary and sentence structures	Students	3.72	.864
	Teachers	3.76	.781
• Providing information or explaining test content related to the CET-4	Students	3.41	.871
	Teachers	3.45	.877
• Organizing classroom activities such as pair work, group discussions	Students	3.21	1.058
	Teachers	3.56	.894
• Explaining learning skills and test-taking strategies	Students	3.19	.900
	Teachers	3.61	.825
• Organizing integrated language activities	Students	2.86	1.154
	Teachers	3.40	1.030
• Organizing language games	Students	2.70	1.129
	Teachers	3.12	1.077

The top three items listed by students were "Explaining the textual meaning" (Mean=4.15), "Explaining textbook exercises" (Mean=4.00) and "Explaining language points" (Mean=3.72). As to teachers' top three rankings, the only difference was that “Explaining textbook exercises” (Mean=3.59) was replaced by “Explaining learning skills and test-taking strategies” (Mean=3.61). Activities like group discussions, pair work were sometimes organized in their classes. According to students’ evaluations, “Organizing integrated language activities” (Mean=2.86) and “Organizing language games” (Mean= 2.70) were seldom practiced in class. Thus, it may be assumed that the normal teaching classes tended to be teacher monologue type and less communicatively oriented.

Next, both students and teachers were asked to evaluate students’ motivations for College English learning on a 5-point Likert scale of agreement. As Table 9.28 displays, “To pursue further studies” and “To further improve English proficiency” received stronger agreement from students, indicating students’ positive attitudes to learn English. It is also understandable for students to top the item “To pursue further studies”.

Table 9.28 Students’ purposes of College English learning

Item	Group	Mean	SD
To pursue further studies	Students	4.00	.933
	Teachers	3.88	.706
To further improve English proficiency	Students	3.92	.941
	Teachers	3.69	.978
To obtain advantage in employment	Students	3.86	1.011
	Teachers	**4.09**	.669
To pass the CET-4	Students	3.75	1.001
	Teachers	3.83	1.020
To satisfy the social needs	Students	3.75	1.018
	Teachers	3.89	.786
To satisfy academic credit requirements	Students	3.52	1.015
	Teachers	*3.97*	.803

In interviews, they explained that today a vast majority of graduates intended to pursue their Master’s degrees home and abroad, which means that they need to take GSEEE, TOEFL or IELTS. Hence, it can be seen that students are clearly aware of the importance to improve their English proficiency. However, teachers gave a very different ranking of these items. They strongly agreed that students' motivations were "to obtain advantage in employment" (Mean=4.09), and "to satisfy academic credit requirements" (Mean=3.97). It is noted that "To pass the CET-4" was not rated as a strong factor motivating students to learn English at university.

It is unknown whether the different rankings were due to teachers' stereotyped perceptions of their students' learning motivations. Apparently, teachers thought their students' learning motivations were instrumental. An examination of students' English study activities and time spent on English interest study may help explain why teachers thought so (see Figure 9.5, Table 9.29).

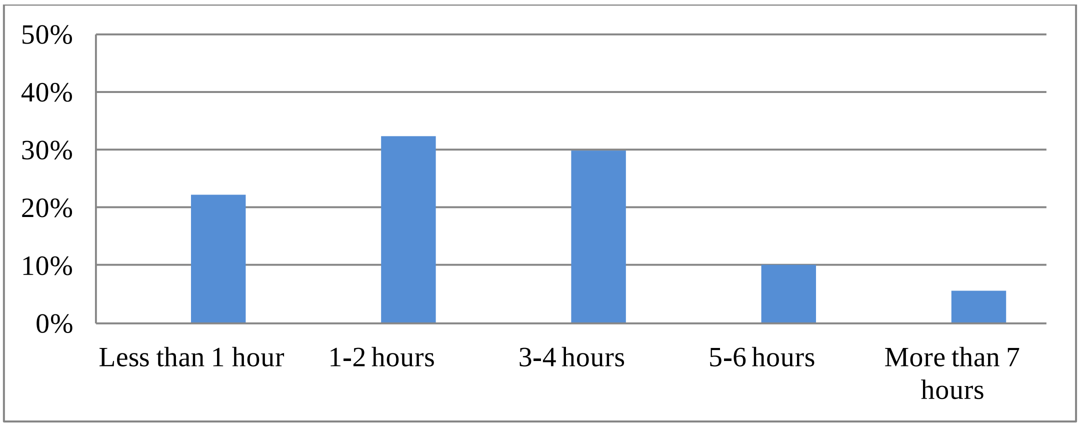

Figure 9.5 Students' average time spent on extra-curriculum English interest study per week

Students reported the average time spent per week on the extra-curriculum English interest study as follows: less than 1 hour (22.2%), %), 1-2 hours (32.4%), 3-4 hours (29.9%), 5-6 hours (10%), more than 7 hours (5.6%). Altogether 54.6% of students spend no more than 2 hours per week learning English for interests.

In addition, students were asked about their English-related activities outside of class. Table 9.29 presents the mean lists on a 5-point Likert scale of frequency. The results revealed that students often did the CET-4 reading test to improve their reading ability (Mean=3.5). They sometimes did the CET-4 listening test, watched films and TV programs in English to improve their listening ability. The mean values of the rest were all below 3.0, indicating students only occasionally read English newspapers, magazines and books, wrote in English, or practiced oral English at English corners and saloons. Thus, it can be seen that outside of class students spent more time on test-related activities rather than on English interest study. They were willing to spend more time on activities to help them pass the CET-4.

Table 9.29 Frequency of students' English-related activities outside of class

Item	Mean	SD
Doing CET-4 reading test to improve my reading ability	3.50	1.003
Doing CET-4 listening test to improve my listening ability	2.97	1.105
Watching films and TV programs in English	2.97	1.146
Reading English newspapers, magazines and books	2.50	1.030
Listening to radio programs in English	1.98	.980
Writing mails and diaries in English	1.83	.964
Practicing oral English at English corners or saloons	1.53	.838

The last category in this section attempted to seek opinions from students and teachers on potential aspects where College English teaching needs further improvements (see Table 9.30). Both groups strongly agreed that students' ability to communicate, esp. their listening ability urgently needed improving. The mean values from teachers basically were above 3.5, indicating teachers thought there was still much room for improvements in all the following aspects, particularly the large class size (Mean=3.81). Teachers reported the typical number of students in their classes as follows: 20-30 (14.1%), 31-50 (30.5%), 51-70 (30.5%), 71-90 (18.8%), 90 or above (6.2%). Over half of teachers had a larger class size containing more than 50 students, which obviously hindered oral activities in class. However, students showed uncertain attitudes to large class size (Mean=2.95) and inadequate teaching hours (Mean=2.97). In terms of the textbooks, both students and teachers shared similar ideas that textbooks and teaching resources should be more diverse. Half of teachers expressed satisfaction with the textbooks they were currently using. Student interviewees also expressed their satisfaction, saying that the textbooks were more interesting than those used in high school.

Table 9.30 Improvements to be made in College English teaching

Item	Group	Mean	SD
Students' ability to communicate, esp. their listening	Students	4.24	.880
ability	Teachers	**4.23**	.825
Inadequate textbooks and teaching resources	Students	3.51	1.047
	Teachers	3.55	1.063
The lack of teaching and learning aids and facilities	Students	3.41	1.002
	Teachers	**3.56**	1.117
Inadequate class hours per week	Students	2.97	1.026
	Teachers	3.42	1.069
Large class size	Students	2.95	1.082
	Teachers	3.81	1.215

9.6 Backing from inferential statistics of questionnaires

In previous sections, descriptive results from both student and teacher questionnaires were reported. The discussions covered the major components in the washback phenomenon, including teachers' and students' perceptions of the CET-4, evaluations of the test design, test preparation activities, test-taking strategies, and teaching and learning behaviors. These findings present a general washabck network in which the CET-4 is related to various aspects in teaching and learning. As noted earlier, literature has proved that learner variables may interact with test characteristics in determining test takers' test performances. However, whether there is a direct relationship between learner variables and CET-4 scores in Chinese EFL testing contexts and to what extent these variables can influence test outcomes are worthy of further investigation. Thus, the following part investigates how students' perceptions of the CET-4 and its washback affect their test performances (RQ3.2). Results from inferential statistical analyses of questionnaires will be reported. One point to make here is that the inferential statistical analyses were not performed on all the questionnaire items but on variables such as students' motivations, factors affecting their test performances and their test taking strategies. As discussed earlier (see section 6.4.4), to achieve the above purposes, several statistical procedures were followed step by step.

First, of the 753 valid student questionnaires, altogether 460 respondents' test scores and their questionnaires were linked correspondingly for inferential analysis. Second, the missing values were replaced by means of items given that no obvious pattern was found. Third, although some items were intentionally grouped together under different subscales, these hypothesized constructs need to be empirically verified. In addition, the number of variables assessed by the questionnaire survey need to be reduced by identifying underlying dimensions and then by computing multi-item scales (Dörnyei, Csizér & Németh, 2006). Thus, exploratory factory analysis was performed first to uncover the latent structure. The principle component analysis was adopted. Oblique rotation was applied because these factors were not assumed to be unrelated or completely independent of each other (Bryman & Cramer, 2012). Multiple regressions were conducted to explore the relationship between the extracted factors and the test performances.

9.6.1 Exploratory factor analysis on students' perceptions

9.6.1.1 EFA on Students' motivations

There has been a great deal of research proving that learning motivations play an important role in learners' test performances. In previous section, I have reported descriptive statistics of students' test-taking motivations and learning purposes respectively (see sections 9.5.1 and 9.5.5). In this section, items from the two categories were combined for analysis. For one thing, the two categories shared similar items related to students' long-term and short-term goals of College English learning and test-taking motivations. For another, research has proved that stakeholders' perceptions of learning and assessment play an important role in their test performances and the assessment in turn influences their learning beliefs and behaviors (Entwistle & Entwistle, 1991). Therefore, EFA was conducted with items from both categories. As discussed earlier (see section 5.2.3.3), principal component analysis was adopted as the extraction method for an empirical summary of data set. Oblique rotation was chose because the underlying constructs are assumed to be correlated.

Altogether ten items from the above two categories underwent EFA and three factors were extracted from the initial results. However, one item (C1.3) was identified with lower and complex loadings across two factors. Given its poor theoretical fit, this item was deleted from further factoring, and the resultant nine items underwent EFA for the second time. The KMO for these items was .684, lower but still acceptable to proceed with EFA. As eigenvalues (see Table 9.31) show, three factors were extracted, whose eigenvalues were higher than one and this three-factor solution accounted for 64.77% of the total variance.

Table 9.31 Factor eigenvalues of students' motivations

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings **(a)**
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total
1	2.788	30.978	30.978	2.369	26.322	26.322	2.394
2	1.999	22.217	53.195	1.509	16.765	43.087	2.085
3	1.042	11.576	64.770	.615	6.830	49.917	2.220
4	.799	8.883	73.653				
5	.704	7.825	81.478				
6	.556	6.183	87.661				
7	.449	4.984	92.645				
8	.370	4.114	96.759				
9	.292	3.241	100.000				

Note. Extraction Method: Principal Component Analysis

Motivation has been defined and classified by different schools of thought. The classical model of language learning motivation is the classification of integrative and instrumental motivations. The former relates to learners' goals to integrate into the target language and culture community while the latter refers to learners' wishes to attain instrumental or utilitarian goals such as meeting school requirements, passing an examination or applying for a job (Gardner, 1985;

Gardner & Lambert, 1972). Later intrinsic-extrinsic model of motivation is developed. Intrinsic motivation refers to a person's aim to bring about certain internally rewarding feelings of competence and self-determination, while extrinsic motivation is carried out in anticipation of a reward from outside, namely, money, prizes, grades or positive feedback (Deci, Connell, & Ryan, 1989).

The focus of this section is not to offer a detailed theoretical and empirical analysis on the broad domain and complex dimensions of motivations, but on the relationship between motivations and test performances. My interpretations of these factors draw on the classical model of integrative and instrumental motivations. In terms of the CET-4, its test uses have been associated with high-stakes decision, which unavoidably made students' motivations more instrumentally oriented. Moreover, studies have suggested that instrumental motivation should receive special attention in EFL context where learners cannot be sufficiently exposed to the target language community (Dörnyei, 1990; Oxford, 1996). Therefore, in developing items exploring students' motivations, I attempted to explore learners' multi-scale of instrumental motivations for learning English and taking the high-stakes test in Chinese higher education context rather than delineate the taxonomy of variables subsumed under the multi-faceted motivations. In the following part, first we will examine whether the pattern matrix produced from EFA can verify the hypothesized scale and identify interpretable clusters of variables.

From Table 9.32, we can see that loadings of all the items on their respective factors range from the .875 to .654. Factor 1 received salient loadings from three items (C1.5, A3.4, C1.6). They concerned the motivations of meeting social demands, and obtaining career advantage. This factor was labeled as employment motivation.

Factor 2 was associated with three items (A3.2, C1.2, A3.3). They were related to passing the CET-4, obtaining merited credits and the Bachelor's degree. This finding was in line with Shi's study (2000) that the major motivation of Chinese undergraduates for English learning was to pass the test and obtain their academic degree. As noted earlier, academic degrees and higher scores have been valued in China's educational and cultural setting. Some researchers proposed that requirement set by educational institutions also affected students' learning motivations (Chen, Warden, & Chang, 2005; Warden & Lin, 2000). Thus, factor 2 was named as academic requirement motivation.

Table 9.32 Pattern matrix of EFA results for the motivation factors

Item	Factor		
	F1	F1	F3
C1.5 I learn college English to obtain advantage in employment.	**.875**	-.020	.061
A3.4 I take the CET-4 to obtain advantage in employment.	**.833**	.083	-.144
C1.6 I learn college English to satisfy the social needs.	**.689**	-.059	.191
A3.2 I take the CET-4 to satisfy academic credit requirement.	-.008	**.874**	.116
C1.2 I learn college English to satisfy academic credit requirement.	-.057	**.810**	.030
A3.3 I take the CET-4 to obtain the bachelor's degree.	.083	**.654**	-.125
C1.1 I learn college English to improve my English level.	-.004	.034	**.841**
A3.1 I take the CET-4 to check my English proficiency.	-.072	-.006	**.770**
C1.4 I learn college English to pursue further studies.	.288	-.033	**.660**

Note. Extraction Method: Principal Component Analysis
Rotation Method: Oblimin with Kaiser Normalization

The three items (C1.1, C1.4, A3.1) loading on factor 3 are pertinent to an individual's self-efficacy beliefs and goals. More specifically, students make efforts to learn English so that they can further improve their proficiency. They take the test to assess their achievement and diagnose their strengths and weaknesses. They expect their efforts to contribute to the future study achievement. Therefore, factor 3 was interpreted as achievement motivation, drawing on the expectancy-value theory (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000).

To understand the general picture of students' learning and test-taking motivation, Table 9.33 summarizes resultant factor names, percentage of variance, means, standard deviations, standardized alpha coefficients, and the correlations for motivation factors. The scale reliabilities were moderately acceptable, all beyond .68. The three factors were not highly correlated. The employment motivation accounted for the largest percentage of the total variance (30.98%), followed by academic requirement (22.22%) and achievement (11.58%) motivations. The mean values of the three composite factors were all around 3.8, revealing students' strong agreement on these factors. It can be concluded that the clustering of items was in accordance with what had been originally hypothesized. Instrumental motivations played a major stimulating role in students' learning motivations.

Table 9.33 Descriptive statistics, standard alpha coefficients, and correlations for the motivation factors

Factors	% of variance	Mean	SD	Std alpha	Correlations		
					F1	F2	F3
F1 Employment	30.978	3.80	.890	.752	1		
F2 Academic requirement	22.217	3.77	.989	.684	.119*	1	
F3 Achievement	11.576	3.75	.783	.698	.317**	-.157**	1

Note. n=460
*p<.05(2-tailed), **p<.01(2-tailed)

9.6.1.2 EFA on students' perceived factors affecting their test performances

In previous section, students were asked about factors affecting their listening, reading, and the overall test performances respectively (see section 9.5.4). In this section, these related items were put together and underwent principal component analysis first to extract any possible composite factors and then to explore to which extent these factors would influence their test performances. Since these factors may increase students' perceived test difficulty and their test performances, they were defined as a general variable of difficulty factors.

The same EFA statistical procedures applied in the last section were repeatedly performed with 17 items from the above three categories. The KMO for these items was .733. As Table 9.34 shows, six factors were extracted, whose eigenvalues were higher than one and this six-factor solution accounted for 60.00% of the total variance.

Table 9.34 Factor eigenvalues of difficulty factors

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings (a)
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total
1	3.636	21.389	21.389	3.636	21.389	21.389	2.297
2	1.681	9.887	31.275	1.681	9.887	31.275	2.391
3	1.395	8.207	39.482	1.395	8.207	39.482	2.222
4	1.224	7.200	46.683	1.224	7.200	46.683	1.804
5	1.180	6.939	53.622	1.180	6.939	53.622	1.521
6	1.084	6.375	59.996	1.084	6.375	59.996	1.521
7	.967	5.691	65.687				
8	.829	4.874	70.561				
9	.799	4.698	75.259				
10	.695	4.087	79.346				
11	.645	3.796	83.142				
12	.597	3.514	86.656				
13	.533	3.135	89.791				
14	.494	2.905	92.696				
15	.440	2.586	95.282				
16	.419	2.464	97.746				
17	.383	2.254	100.000				

Note. Extraction Method: Principal Component Analysis
Rotation Method: Oblimin with Kaiser Normalization

From Table 9.35, we can see that loadings of all the items on their respective factors ranged from the .862 to .487. Factor 1 was related to three items (A15.3, A19.3, and A16.3) about background knowledge. Factor 2 had loadings from four items (A15.1, A16.1, A15.2, A16.2) which all concerned vocabulary and sentence structure. This factor was labeled as linguistic knowledge. Factor 3 received loadings from four items (A16.5, A15.5, A16.4, A19.4), which involved listening, reading, and test-taking strategies. Factor 4 received loadings from two items (A19.2, A19.1) related to language difficulty. Factor 5 had two items (A15.7, A16.6) covering distractor designing. Factor 6 had two items (A19.5, A19.6) related to time insufficiency and test anxiety.

Table 9.35 Pattern matrix of EFA results for students’ perceived difficulty factors

Items	Factor loadings					
	1	2	3	4	5	6
A19.3 My overall CET-4 performance is affected by unfamiliarity with topics.	**.777**	.078	-.073	.189	-.040	-.004
A15.3 My listening performance is affected by lack of background knowledge.	**.767**	-.095	.009	-.049	.078	-.013
A16.3 My reading performance is affected by lack of background knowledge.	**.721**	-.188	.116	-.097	-.065	-.035
A15.1 My listening performance is affected by limited vocabulary.	-.060	**-.743**	-.015	-.023	.093	-.070
A16.1 My reading performance is affected by limited vocabulary.	.041	**-.725**	.052	.117	-.118	-.037
A15.2 My listening performance is affected by long and complex sentences.	.152	**-.613**	-.059	-.055	.192	.173
A16.2 My reading performance is affected by long and complex sentences.	.161	**-.596**	.086	.110	-.025	.070
A16.5 My listening performance is affected by lack of reading skills.	.016	-.015	**.847**	-.001	-.035	-.045
A15.8 My listening performance is affected by lack of listening skills.	.089	.070	**.687**	.040	.234	-.205
A16.4 My reading performance is affected by slow reading speed.	-.139	-.215	**.561**	.028	.018	.218
A19.4 My overall CET-4 performance is affected by lack of test-taking skills.	.177	.131	**.487**	.023	-.158	.425
A19.2 My overall CET-4 performance is affected by difficulty of questions.	.011	.094	.006	**.862**	.078	.077
A19.1 My overall CET-4 performance is affected by difficulty of language.	.007	-.184	.026	**.787**	-.044	-.066
A15.7 My listening performance is affected by tricky options.	-.006	-.018	.091	-.122	**.808**	.028
A16.6 My reading performance is affected by tricky options.	-.028	-.097	.011	.207	**.647**	-.026
A19.5 My overall CET-4 performance is affected by time insufficiency.	-.150	-.176	-.013	.029	-.055	**.852**
A19.6 My overall CET-4 performance is affected by test anxiety.	.235	.194	-.027	.015	.271	**.493**

Note. Extraction Method: Principal Component Analysis
Rotation Method: Oblimin with Kaiser Normalization

After an examination of scale reliabilities, factor 5 (.404) and factor 6 (.224) were excluded from multiple regression analysis given their lower reliabilities. Table 9.36 summarizes the resultant factor names, percentage of variance, means, standard deviations, standardized alpha, and the correlations for the difficulty factors.

Table 9.36 Descriptive statistics, standard alpha coefficients, and correlations for the difficulty factors

Factors	% of	Mean	SD	Std	Correlations			
	variance			alpha	F1	F2	F3	F4
F1 Background knowledge	21.389	3.38	1.018	.678	1			
F2 Linguistic knowledge	9.887	3.76	.902	.685	-.116*	1		
F3 Skills & strategies	8.207	3.53	.993	.636	.210**	-.172**	1	
F4 Language difficulty	7.200	3.89	.975	.648	.148**	-.197**	.125**	1

Note. n=460
*p<.05(2-tailed), **p<.01(2-tailed)

Factor 1 accounted for the largest percentage of the total variance (21.98%), while the other three factors contributed similar percentage to the total variance.

Students agreed that language difficulty (Mean=3.89) and linguistic knowledge (Mean=3.76) were the two influencing factors on their test performances. Multiple regression analyses conducted in the following part would help confirm students' perceived difficulty factors.

9.6.2 Multiple regression analysis on students' perceptions

Multiple regression analysis is to examine the effect of multiple independent variables on only one dependent variable. Thus, it was applied to answer the research question (RQ3.2) on how students' perceptions affect their test performances. Principal component analyses had extracted three motivation composite factors and four composite difficulty factors. Therefore, the seven composite factors would serve as independent variables to explain their relationship with test takers' CET-4 performance that was treated as dependent variable. Stepwise regression was adopted in this study. It combines both forward selection and backward elimination. When an added variable contributes to the model, it will be kept. Then the rest of variables in the model are assessed. If they fail to contribute significantly, they are removed. One advantage of the stepwise method is that it can result in the minimum number of predictor variables (Tabachnick & Fidell, 2001).

Usually a simplified regression output table is presented to summarize the results. However, given it was the first regression analysis conducted in this study, a few more tables and detailed explanations to important columns were presented here to facilitate our understanding of the following analyses.

The results of stepwise multiple regression analysis revealed that five factors emerged as significant predicators of CET-4 performances (see Table 9.37). As seen in the R square column, Model 1 with just achievement motivation factor accounted for 3.8% of the variance. From the R square Change column, we can see the linguistic knowledge factor independently added 2.8% of explanation for the variation of CET-4 total scores. In the same token, it can be found that the model with five explanatory variables could explain 8.8% of the variance of CET-4 students' overall performances.

In the stepwise regression Model Summary, the F Change Column and the Sig. F Change column revealed the results of an ANOVA, comparing the current model with the previous model to see whether the two models were statistically different. For example, Model 2 was statistically different from Model 1 (F=13.86, *p*=.000), while Model 5 was not statistically different from Model 4 (F=2.90, *p*=0.089 > *p*=.05).

Table 9.37 Model summary of stepwise regression on students' perceptions of the CET-4

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.196[a]	.038	.036	54.883	.038	18.218	1	458	.000	
2	.258[b]	.067	.062	54.128	.028	13.856	1	457	.000	
3	.273[c]	.074	.068	53.963	.008	3.810	1	456	.052	
4	.287[d]	.082	.074	53.785	.008	4.014	1	455	.046	
5	.297[e]	.088	.078	53.673	.006	2.899	1	454	.089	.182

Note. a. Predictors: (Constant), Achievement

b. Predictors: (Constant), Achievement, Linguistic knowledge

c. Predictors: (Constant), Achievement, Linguistic knowledge, Language difficulty

d. Predictors: (Constant), Achievement, Linguistic knowledge, Language difficulty, Requirement

e. Predictors: (Constant), Achievement, Linguistic knowledge, Language difficulty, Requirement, strategies

f. Dependent Variable: CET total scores

Table 9.38 displays coefficients of regression results, revealing information about each explanatory variable in the equation. The unstandardized coefficients help to write each regression model. For instance, in Model 2 factors predicting CET-4 scores could be modeled by the equation below:

$$Y= 460.274+ (11.195)*\text{achievement motivation} + (9.408)*\text{linguistic knowledge}$$

Table 9.38 Coefficients of the stepwise regression on students' perceptions of the CET-4

Coefficients[a]											
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	460.274	2.559		179.871	.000					
	Achievement	10.934	2.562	.196	4.268	.000	.196	.196	.196	1.000	1.000
2	(Constant)	460.274	2.524		182.378	.000					
	Achievement	11.195	2.527	.200	4.430	.000	.196	.203	.200	.999	1.001
	Linguistic knowledge	9.408	2.527	.168	3.722	.000	.163	.172	.168	.999	1.001
3	(Constant)	460.274	2.516		182.938	.000					
	Achievement	11.011	2.522	.197	4.367	.000	.196	.200	.197	.998	1.002
	Linguistic knowledge	8.415	2.571	.151	3.274	.001	.163	.152	.148	.960	1.042
	Language difficulty	-5.018	2.571	-.090	-1.952	.052	-.126	-.091	-.088	.960	1.042
4	(Constant)	460.274	2.508		183.541	.000					
	Achievement	11.815	2.545	.211	4.642	.000	.196	.213	.208	.973	1.028
	Linguistic knowledge	8.812	2.570	.158	3.429	.001	.163	.159	.154	.954	1.048
	Language difficulty	-5.505	2.574	-.098	-2.139	.033	-.126	-.100	-.096	.951	1.051
	Requirement	5.140	2.566	.092	2.004	.046	.033	.094	.090	.957	1.044
5	(Constant)	460.274	2.503		183.924	.000					
	Achievement	11.783	2.540	.211	4.639	.000	.196	.213	.208	.973	1.028
	Linguistic knowledge	8.174	2.592	.146	3.154	.002	.163	.146	.141	.934	1.070
	Language difficulty	-5.141	2.577	-.092	-1.995	.047	-.126	-.093	-.089	.945	1.058
	Requirement	5.622	2.576	.101	2.183	.030	.033	.102	.098	.946	1.057
	strategies	-4.377	2.571	-.078	-1.703	.089	-.106	-.080	-.076	.950	1.053

Note. a. Dependent Variable: CETOTAL= CET total score.

The t-test column in the Coefficients table tests whether each explanatory variable contributes uniquely to the equation (Tabachnick & Fidell, 2001). Whether each coefficient is statistically significant can be seen from the Sig. column and stars will be correspondingly marked on them in their equation. The last column indicates information about multicollinearity. The VIF values of over 5 are evidence of collinearity, indicating the variables are too highly correlated and may harm the model (Heiberger & Holland, 2004, p. 243). The VIF for this data set was around 1, so no variance inflation factor was found.

Three charts generated from output helped examine regression assumptions of normality of data, outliers and homogeneity of variances. In order to examine the normality of the data, the distribution of the residuals rather than the distribution of the individual variables should be examined. The P-P plot of the standardized residuals shows a normal curve (see Figure 9.6). No extreme variables could be identified as evidence of non-normality of data distribution in Figure 9.7. Figure 9.8 pictures a randomly scattered scatterplot, confirming that the assumption of homoscedasticity was met.

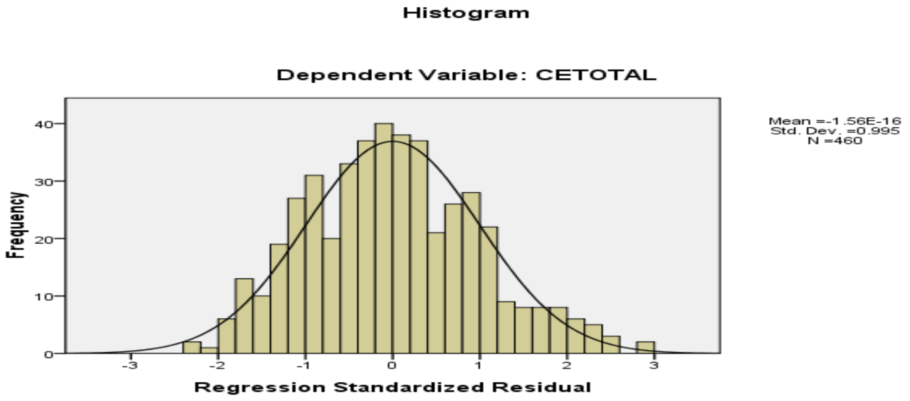

Figure 9. 6 P-P plot of the regression standardized residuals

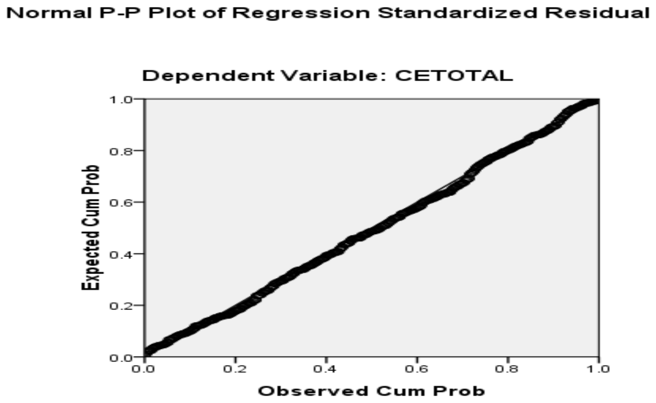

Figure 9.7 P-P plot for diagnosing normal distribution data

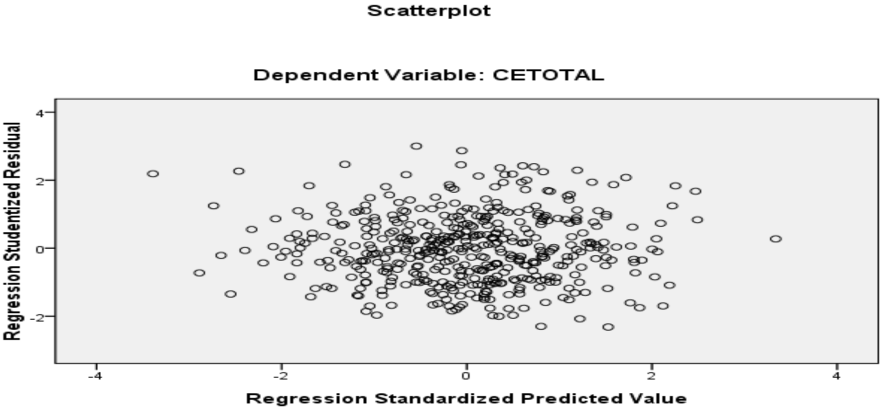

Figure 9.8 Plot of studentized residuals crossed with fitted values.

To sum up, prior to reporting the results, the assumptions of multiple regressions should be examined as discussed in the above part. When writing the results, a simple table can be produced but includes the most important statistics like Beta, t values, total R square for the model, and unstandardized regression coefficients. In the above stepwise regression, five factors emerged as significant predicators of CET-4 performances, accounting for 8.8% of the total variance of CET-4 scores. Students' achievement motivation emerged as the strongest predictor, accounting for 3.8% of the variance of CET-4 performances. The second strongest factor was linguistics knowledge, accounting for an additional 2.8% of the variance in CET-4 scores, followed by factors of language difficulty, academic requirement motivation and test-taking strategies. However, the last three factors accounted for only about 1% of variance of CET-4 performances. The regression equation for model 5 was presented below. It can be noticed that the equation contained two negative unstandardized coefficients, indicating that students' CET-4 scores would be decreased with the increasing difficulty of vocabulary and sentence structure in test content and students' failure to apply test-taking strategies to handle these difficulties.

Y= 460.274+ (11.783)*achievement motivation + (8.174)*linguistic knowledge + (-5.141) language difficulty + (5.622)*academic requirement motivation + (-4.377) strategies

The above findings have practical implications. Those who reported to learn English and take the CET-4 for further improvement achieved higher scores than those who are instrumentally motivated to get employment advantage. Thus, students should adjust their learning purposes and test-taking motivations to be more integrative or achievement oriented motivated. With regard to difficulty factors, vocabulary knowledge and complex sentence structure have more influences on their scores. Students' background knowledge does not constitute as a predictor, which may be taken as supportive evidence that test content does not favor a particular group of test takers.

9.6.3 Multiple regressions on students' test-taking strategies and their performances

This section attempts to explore effect of students' test-taking strategies on their CET-4 performances. As discussed earlier, the study is not to specifically target at the relationship between learner strategies and their scores. Therefore, the student questionnaire did not cover the taxonomy of strategies but mainly drew on those listed in the CET-4 syllabus. In addition, in the preliminary, a few strategies frequently employed by students and those recommended by teachers were identified. Then they were included in the questionnaire for examination. EFA failed to extract interpretable factors given the complex loadings across over two factors and poor theoretical fit. Therefore, these items directly underwent stepwise regressions. Five listening strategies were treated as independent variables and CET-4 listening scores as the dependent variable. One more stepwise regression was conducted with eight reading strategies as independent variables and CET-4 reading scores as the dependent variable. The same regression procedures discussed in last section were repeated in this section. The assumptions of normality, outliers, and homoscedasticity were confirmed and no violation was identified. Tables 9.39 and 9.40 summarize the major findings respectively. The relationship between listening strategies and listening scores were examined first (see Table 9.39).

The results revealed that two strategies emerged as significant predictors of CET-4 listening performances. In this model, the multiple correlation coefficient of the two factors is .278, accounting for 7.7% of the variance in CET-4 listening scores (*p*<.01). Factor of "taking notes to help memory" emerged as the strongest predictor of listening performance, making significant contribution (*β*=.251, *p*<.01). The strategy of "catching conjunctions to infer speakers' attitudes" accounted for 1.4% of the variance in listening scores (*β*=.124, *p*<.01). Strategies like "reading options first for prediction", "skipping unknown words", "inferring speakers' opinions from stress and intonation" were excluded as non-significant factors to predict students' CET-4 listening performances.

Table 9.39 Summary table of regression results: students' listening strategies and their listening performances

	Factor	R	R^2	ΔR	*F change*	F	B	*β*	t
1	A17.2 Taking notes to help memory	.251	.063**	.063	30.786	30.786	5.879	.251	5.549
2	A17.2 Taking notes to help memory	.278	.077**	.014	7.100	19.148	5.110	.218	4.681
	A17.5 Catching conjunctions to infer speakers' attitudes						3.251	.124	2.655

Note. ***p*<.01

Table 9.40 summarizes regression results on students' reading strategies and CET-4 reading performances. This model accounted for 5.7% of the variance of students' CET-4 reading performances (*p*<.01). Among the three predictors, the strategy of "guessing unknown words from the context" served as the strongest predictor, accounting for 3.9% of the variance of reading scores. Strategies of "analyzing grammatical structure of complex sentences" and "scanning for specific details" added the prediction power in this model by 1.2% and 0.6% respectively. The rest of strategies such as "predicting the content by options, skimming for the main idea and skipping the unknown words" were not identified ad significant factors to predict students' reading performances.

Table 9. 40 Summary table of regression results: students' reading strategies and their reading performances

	Factor	R	R^2	ΔR	*F change*	F	B	*β*	t
1	A235 Guessing unknown words from the context	.197	.039**	.039	18.447	18.447	5.116	.197	4.295
2	A235 Guessing unknown words from the context	.225	.051**	.012	5.765	12.201	4.493	.173	3.704
	A233 Analyzing grammatical structure of complex sentence						2.460	.112	2.401
3	A235 Guessing unknown words from the context	.238	.057**	.006	2.957	9.155	3.917	.151	3.118
	A233 Analyzing grammatical structure of complex sentence						2.191	.100	2.118
	A238 Scanning for specific details						2.137	.083	1.720

Note. **p<.01

To sum up, questionnaire results revealed that strategies such as "reading options first" and "skipping the unknown words" tended to be frequently adopted by students in their test-taking processes (see section 9.5.4). However, results from multiple regressions proved that these strategies actually had no effect to predict their listening performances. Students showed least agreement to "taking notes to help memory", but it turned out that those who took notes to help their memory tended to get higher scores.

Likewise, strategies that students reported to frequently apply in taking reading test did not contribute to the improvement of their reading scores. However, it should be noted that those strong predictors on listening and reading scores were all topped in teachers' ranking list, indicating teachers' emphasis on and encouragement of employing these effective strategies. It is expected that the above findings would be informative to students so that they can raise their awareness of more effective strategies.

9.7 Summary

This chapter discussed backing evidence related to the consequence claim and sought answers to the two sub-research questions (RQ3.1, RQ3.2). Evidence generated from both qualitative and quantitative data helped to delineate stakeholders' perceptions, attitudinal and behavioral patterns pertaining to the CET-4.

The first part of the chapter was related to how stakeholders perceived the CET-4 and its washback. Specific columns in teacher questionnaire helped explore their understandings of background behind the 2006 CET-4 reform, their perceived changes in terms of the revised test design, and their assumed theoretical rationales underlying these changes. Given the common themes in both questionnaires, students' and teachers' responses to the overlapping items were grouped together and discussed in a comparative approach. The aspects under investigation mainly covered stakeholders' perceptions of test design, test-taking and preparation activities, and their normal teaching and learning behaviors. Similarities and differences were identified and the possible reasons leading to these discrepancies were also explored.

The second part of this chapter explored the relationship between students' perceptions and their test performances. First, principal component analysis helped to extract interpretable composite factors. Next, these factors along with the test performances underwent multiple regression analyses. The second part narrowed down the focus on students' motivations, perceived factors affecting their listening and reading comprehension, and the strategies they preferred in taking listening and reading comprehension.

CHAPTER 10

CONCLUSIONS

10.1 Introduction

Following a wide range of backing evidence discussed in previous chapters, this chapter will synthesize findings obtained from multiple sources to revisit the research questions and their corresponding claims in the AUA for the CET-4. Then implications generated from findings are addressed. The chapter concludes with limitations of the study and suggestions for future research.

10.2 Summary of main findings

In this section findings from both preliminary and main studies will be synthesized. It will start with a revisit to the overall application of an AUA, since building a specific AUA for the CET-4 is the primary step guiding the study to seek answers to research questions. It will then summarize main findings in relation to research questions along with their corresponding claims.

10.2.1 The AUA for the CET-4 revisited

A distinctive feature of this study is the articulation of an AUA for the CET-4 within China’s higher education assessment context. The study built an AUA to investigate the use of the CET-4 in instructional setting while weighing the construct validity of its revised listening and reading components. The originality and significance of this study owed much to application of Bachman and Palmers’ AUA. First, the AUA provided a conceptual framework to guide the whole study, and a series of feasible procedures to develop a test or justify its uses. Second, the AUA enabled the study to show a due concern about decisions made based on CET-4 scores. Compared with other argument-based approaches to validation, the claim on decisions not only distinguished decisions that are made based on test scores from consequences of test uses, but also served as an essential bridge linking claims on score interpretations and test consequences together. Above all, via a series of logical and coherent inferences, the AUA provided a link to discuss washback issues in relation to validity.

The study in return has contributed its value to this conceptual framework by offering an exemplary attempt to justify the legitimacy of applying an AUA to evaluate an existing assessment, and by demonstrating how to adapt a full-scale AUA template to the specific CET-4 context with necessary modifications and flexibility. The study has also contributed to the ongoing development of the CET-4 by addressing both validity and washback issues.

However, it is noteworthy that application of the AUA framework poses higher requirements for data availability. When drawing on the AUA to justify or evaluate an assessment, the research may be constrained in its scope and depth due to limited access to data sources, particularly test data and test specifications. This also explained why the present study only concentrated on three claims.

Drawing on the AUA, the study sought answers to three research questions. In the following part, the three questions and their sub-questions will be revisited to highlight major findings and propose some thought-provoking issues in relation to the CET-4.

RQ1: To what extent can the CET-4 serve as an indicator of students' English proficiency?

RQ2: What evidence has been provided or is needed to justify the major types of decisions made based on CET-4 scores?

RQ3: In what ways and to what extent can the CET-4 and the decisions made based on it affect English teaching and learning?

10.2.2 The interpretation claim and RQ1

The interpretation claim sought answers to the first research question and two sub-questions. This part of study examined construct of the CET-4 and its content representativeness and coverage in alignment with the CECR and the CET-4 Syllabus. Analyses of test scores and test papers provided major backing evidence for this claim. Documents like teaching and testing syllabuses, and the scholarly articles by NCETC served as references in interpretation of test scores as well as benchmark in content analysis.

Findings on test construct

RQ 1.1 To what extent does the CET-4 measure the construct to be assessed?

In the preliminary study, a statistical comparative study between the old and the new versions of the CET-4 was conducted with 188 valid sets of test data cases to explore the internal structure of the test. Correlations between the total scores and the revised listening and reading components in the new CET-4 (.867, .834) were higher than their counterparts in the old version (.777, .785). The intercorrelations and the shared variances revealed the proportion of the construct overlap of the corresponding components in the two tests. The two tests (56.9%) especially their Listening components (49.1%) shared about half of the same construct. This finding was in line with the original intentions of 2006 CET-4 reform, adding new elements to further improve the validity of CET-4 while retaining its original merits. The lowest value of .132 indicated that there was only 13.2% overlap of the variance between the scores on the reading components of the two tests. In other words, 86.8% of the two reading subtests measured different skills. It can be explained that the reading component of the new CET-4 did undergo a dramatic modification both in its construct and test methods. As mentioned earlier, the sections of Skimming and Scanning and Banked Cloze in the post-2006 CET-4 have replaced two of the careful reading passages in the pre-2006 CET-4. Apparently, skills used in the careful reading and fast reading are different. Moreover, banked cloze requires better discourse skills and lexical knowledge than just reading for gist or details. Principal component analysis showed that only one factor could be extracted from each test. However, the interpretations of the factor should be different in light of the incentives and purposes of the CET-4 reform and different factor loadings in the two tests. In other words, the language abilities to be measured, hence the constructs of the two tests should have distinctions. The highest factor loading (.837) on the listening component of the new CET-4 was in line with the underlying rationale and design of the 2006 CET-4, laying more emphasis on measuring students' listening proficiency.

In the main study similar statistical procedures were performed on a larger pool of 2692 valid post-2006 CET-4 score cases (see Chapter 7). Findings from both preliminary and main studies were cross validated, indicating that the current CET-4 possessed a better structure in sense of correlation analyses. The results proved that listening, reading, integrative and writing skills were all examined in the current CET-4, which in turn contributed to one general factor.

To make score-based interpretations more meaningful, the test construct was addressed with reference to relevant documents including the uniform CECR, the CET-4 Syllabus and publications of the NCETC. The NCETC defined the construct to be an assessment of comprehensive employment of English or overall English proficiency (see section 7.2). Only in one article by Jin (2008), Chair of NCETC, the construct was labeled as the Communicative Language Ability.

It seemed that the NCETC hedged to delineate a well-defined construct definition. Therefore, this study recommends the NCETC addressing test construct in detailed and professional technical terms whatever it is based on, a needs analysis, an instructional syllabus, or a language ability model. For example, the NCETC declares that the CET-4 aims to measure students' overall English ability and to maximize its positive effect on College English teaching. They can be more specific on whether the overall English proficiency refers to the widely acknowledged Communicative Language Ability, or what beneficial consequences the test is expected to promote. Likewise, I would also like to draw attention to the long contention on whether the CET-4 should be designed as an achievement test or a proficiency test. The NCETC maintains that the CET-4 has been designed as an achievement test since its administration, because the test, designed in line with the teaching syllabus, aims to examine the teaching and learning outcome of foundation stage and serve College English teaching. However, some scholars argue that the CET-4 is a proficiency test, because almost all the university students are required to take the same test, regardless of various textbooks in use and different teaching curriculums in implementation at different universities. Whether the NCETC maintains its assertion or acknowledges the CET-4 as a proficiency test, the NCETC should provide backing evidence.

The argument made here is that the NCETC should not expect everyone to buy the construct label they stick to the test or accept all their claims. Instead, they should offer more accurate, detailed and explicit definition on the test construct and related backing evidence to support the credibility of their claims and to refute rebuttals. Just as Bachman and Palmer (2010) stress, specific definitions of constructs provide theoretical underpinnings for both test development and justification of the intended assessment-based interpretations.

Findings on content representativeness and coverage

RQ1.2 To what extent is the CET-4 representative of the content relevance and coverage in accordance with the test syllabus and curriculum objectives?

Based on a modified framework of task characteristics proposed by Bachman and Palmer (1996, 2010), content analyses were conducted with listening and reading tasks in seven test papers, based on six parameters of text length, readability, topics, genres, listening and reading skills coverage. In addition, a diachronic approach was adopted as well to compare some of the findings from the present study with those from similar studies conducted in the first two stages of CET-4 development. Such comparisons helped reveal changes and shifting focus embodied from the test design.

The average length of both listening and reading comprehension tasks are within the length range proposed by the CET-4 Syllabus. The changing trends in readabilities of listening and reading components in the past years are in line with the changing focus of the post-2006 CET-4 shifting from its traditional emphasis on examining students' reading ability to examining students' listening ability. Genres specified in both the CETCR and the CET-4 Syllabus are covered by certain proportions of passages. Narration occupies the largest proportion in listening passages, followed by argumentation. Exposition and argumentation are the frequently used genres in reading passages. In terms of topics, short conversations are related to students' daily life and academic study, while long conversations cover scenarios of job-hunting, interviews, hotel checking, etc, where students are highly likely to be exposed to in their future work and life. The listening conversation parts demonstrate higher degree of authenticity, which is also confirmed by student in their interviews. Social science is the mostly adopted topic in listening and reading passages, followed by natural science and humanities. Almost all the skills listed in the CECR and the CET-4 Syllabus are covered in the seven test papers. Short conversations lay more emphasis on examining the skill of *making inferences and deductions*. The skill of *understanding important and specific details* is attached great importance in both long conversations and short listening passages. In reading component, the skill of *scanning for locating specific information* is fully examined in fast reading while skills of *understanding both explicitly and implicitly stated conceptions or details* are frequently tested in careful reading. In brief, results from test content analyses have demonstrated an overall neat correspondence between what test tasks are designed to assess and what the teaching and testing syllabuses stipulate to assess.

However, a few points are worth discussion here. First, it is noted that listening and reading skills stipulated in the 2006 CET-4 Syllabus showed no differences with those in previous syllabus. Since the CET-4 underwent the largest reform in both test content and format, inclusion of long conversations, compound dictation in the listening component, fast reading and banked cloze in the reading component would bring about shifting focus on language abilities or skills that the test emphasizes. In addition, statistical analyses of test scores revealed that half of the test constructs, especially the listening construct were changed. It follows that the intended skills to be measured may undergo some adjustments accordingly. However, no changes in skills listed in CET-4 Syllabus can be identified.

Second, take listening skills coverage for example, questionnaire findings revealed heavy examination on skill of "*making inferences and deductions*" in short conversations. However, skills of "recognizing phonological features and communicative functions of utterances" appeared to have little or no coverage. It may be argued that "recognizing phonological features" is a basic skill for listening comprehension, so it is unnecessary to design specific items to examine this skill. It is advised to replace such basic skills with either advanced or frequently adopted skills. The argument here is that those rarely tested skills deserve test designers' attention, because they may represent construct underrepresentation or construct-irrelevance variance. Both would threaten the test construct. Thus, the study recommends that the skills listed in CET-4 Syllabus should be more specific and detailed, or be stipulated with reference to detailed taxonomies of sub-skills generated from previous research (e.g., Alderson, 2000; Buck, 2001).

Third, the TLU domain of the CET-4 is not explicitly delineated. As discussed earlier, the teaching objective is to develop students' ability to use English in an all-round way so that they can communicate effectively in their future work and social interactions. In this sense, it is understandable that the TLU domain is too broad to be generalized. However, this may constitute a potential rebuttal to the generalizability of score interpretations. Since the CET-4 is intended to examine the teaching effect of College English teaching at the foundation stage, the TLU domain of the CET-4 is advised to be defined as academic English use at campus. A clear delineation of TLU is also closely related to definition of test construct.

In brief, based on data available in this study, it can be concluded that students' CET-4 performances can be interpreted as a useful indicator of their English proficiency. The interpretations of scores are meaningful in terms of construct defined in the teaching curriculum and testing syllabus, and are impartial to all the test takers due to a series of procedures ensuring its score reliability and bias free content. The test demonstrates good internal structure of construct and tasks in the test papers basically cover the intended skills defined in the syllabus.

10.2.3 The decision claim and RQ2

The decision claim serves as an important inferential link between score interpretations and test consequences. It was articulated in response to the second research question. The warrants related to value-sensitivity and equitability guide the investigation into factors underlying the decision-making process.

The study has described decisions made on CET-4 scores at two levels. The nationwide decisions made by the NCETC are related to test takers' eligibility to obtain the Score Report Form, and take the CET-4 SET and the CET-6. Questionnaire findings have revealed that teachers (64.1%) showed more positive attitudes than students (35.4%) to the SRF. Both teachers and students agreed composite and profile scores reported on the SRF enabled students to better diagnose their strengths and weaknesses in language skills. However, students complained in interviews that the SRF exerted more pressure on them since they had to score higher to impress the future employers when the specific scores rather than a score scale were reported. In addition, it is found that students' major dissatisfaction resulted from their difficulty in understanding the standardized score transformation. Therefore, the study suggested that verbal descriptions of proficiency levels should be established in accordance with each score range so as to make score interpretations more meaningful.

Decisions made at the institutional levels were summarized from interviews and questionnaire findings. Universities tended to tie students' CET-4 performances to decisions related to graduation, placement, and curriculum adjustment. The cut score of 425 is set as the threshold in making these decisions. It is found that these decisions tended to induce strong criticisms and disputes. Therefore, the study, under the guidance of warrants, explored reasons behind these decisions from the perspectives of value-sensitivity and equitability.

The score-based decisions are found to take a deep root in the inherent imperial examination system in China and the long-standing societal and cultural values. Tests have been long accepted as a fair tool in China for making decisions in selection, competition and advancement. Better test performances tend to be acknowledged as an indicator of success. Thus, the importance attached to testing has nurtured the increasing power and the high-stakes of large-scale tests. Another contributing reason is that China has a long history of centralized, top-down and examination-based educational system. Decision-makers are likely to impose unintended uses or even misuses on the test. Take the graduation decision for example, the assumed scenario by administrators of the University Academic Affairs Office was that the high-stakes of graduation decision would make students more accountable for their study, and promote the effective teaching and learning in their universities. However, even though this policy stimulated students to take their English learning seriously, it also subjected universities to criticisms or even lawsuits, because no regulations and legal documents could be traced to support this decision.

Therefore, the study makes an argument that it may be sufficient and relevant to interpret CET-4 scores below 425 as a student’s failure to reach the required English proficiency. However, it is only relevant but insufficient enough to make a decision that this student fails in his major learning and in his university life so that he cannot be awarded his Bachelor’s degree. The decision to tie CET-4 scores with academic degrees is a good case in point, embodying universities’ central power and top-down policies. As Bachman and Palmer suggest (2010), both test developers and test users need to consult the immediate and direct stakeholders, namely, test takers, as well as other groups of stakeholders in their decision-making process. They should also take into account societal values like fairness and ethics to assure the appropriateness of their decisions.

Another indication generated from this claim is that test users should be responsible for decisions they have made and held accountable to stakeholders to be affected by their decisions. However, it does not mean test developers are free from accountability. They should improve test qualities so that decisions can be made based on valid, reliable, meaningful, and generalizable interpretations of scores.

Hence, this study raises a question for consideration: whether test agencies should guide test users in making proper decisions based on test scores, and what guidance and suggestions test agencies can offer to hinder misuses of a test. With regard to the CET-4 and the graduation decisions, it is not enough for the NCETC just to reiterate that they have never proposed this requirement, and they do not support this practice. It deserves a concern on what the NCETC can do to guide universities' uses of scores.

10.2.4 The consequence claim and RQ3

The consequence claim in an AUA was articulated in accordance with the third research question. Three warrants centered on the quality of beneficence. Since washback is a complex mechanism subsuming multi-faceted dimensions, the focus was narrowed down to investigate stakeholders' perceptions of the CET-4 and its washback and the relationship between their perceptions and their CET-4 test performances. Questionnaires and interviews served as dominant instrument in exploring teachers' and students' perceptions of test design, test influences, test preparation and test-taking activities, as well as English teaching and learning practices.

Findings on stakeholders' perceptions of the CET-4

RQ 3.1 How do stakeholders perceive the CET-4 and its washback?

Stakeholders' perceptions of test design were more positively oriented. Both teachers and students believed that the CET-4 served as an effective means of measurement. It provided a relatively objective, reliable and standardized evaluation of the teaching and learning outcomes. Teachers evaluated test tasks as more communicatively oriented and laying more emphasis on developing listening ability, which was in line with the intended purposes of the 2006 CET-4 reform. Students evaluated listening component as the most communicatively oriented, and compound dictation as the most difficult task.

In terms of the overall influence of the CET-4, students and teachers showed some divergences. Questionnaire analyses revealed that teachers believed the long-standing influences of the CET-4 were more negatively oriented. During the normal teaching period, the washback of the CET-4 on classroom activities is not obviously discovered. However, with the CET-4 approaching, the negative effects can be more intense, because its high-stakes nature induced the phenomenon of "teaching and learning to the test". The typical practice was to stop normal teaching and start test preparation courses. The positive influence is that the CET-4 has urged universities to attach great importance to College English teaching and learning. It has made great contribution to improving the social status and quality of College English teaching. When students were asked to evaluate the influences of the CET-4, they believed that its overall influence was more beneficial. In fact, students were found to hold complicated and mixed feelings to this test. On the one hand, the CET-4 brought about pressure to them with its high-stakes decisions and far-reaching influences. Some students expressed strong discontent with the hidden policy to link CET-4 performances with their Bachelor's degrees. On the other hand, some students frankly admitted that they would not be motivated to learn English if they were not required to take the

CET-4, or if the test results were not linked to the graduation decision. The CET-4 stimulated them to take their English learning seriously.

Regarding test preparation activities, students believed that preparation for the CET-4 would improve their performances, but they did not think their teachers' test preparation courses helped significantly. The test preparation methods students advocated were to memorize vocabulary and do mock test papers. In terms of test-taking activities, students rated "complex sentence structures", "limited vocabulary" and "tricky options" as the top factors influencing their listening and reading performances. With respect to test-taking strategies, students laid more emphasis on "reading options first for predication" and "guessing or skipping unknown words". However, multiple regression analyses revealed that "taking notes to help memory" and "catching conjunction to infer speakers' attitudes" were effective strategies to improve listening scores.

As to College English teaching and learning, teachers evaluated teaching experience and teaching belief as the top two factors affecting their English teaching. When students were asked to evaluate their teachers' classroom activities, they rated explaining the textual meaning, textbook exercises, and language points as the top.

Findings on relationship between students' perceptions of the CET-4 and their test performances

RQ 3.2 How do students' perceptions affect their test performances?

Principal component analysis and multiple regression analysis were conducted with some questionnaire items to explore possible relationships between students' perceptions of the test and their test performances.

First, a three-factor motivation solution was generated by PCA. Employment motivation accounted for the largest percentage of the total variance, followed by academic requirement motivation and achievement motivation. In terms of difficulty factors affecting students’ test performances, four factors were extracted and interpreted as factors of background knowledge, linguistic knowledge, test-taking skill and strategies, and language difficulty.

Second, the three motivation composite factors and four composite difficulty factors altogether served as independent in the stepwise regression analysis, with the CET-4 performances treated as dependent factors. The results revealed that five factors emerged as significant predicators and could explain 8.8% of the variance of students’ CET-4 performances. Achievement motivation and linguistic knowledge were the top two predicators.

10.3 Implications of the study

Based on the findings generated from the present study, some implications can be made from the study.

There has been a long debate in language assessment filed on who should be responsible for the negative consequences of an assessment. Multiple uses of the same assessment present challenges to the assessment development, particularly when an assessment is used for unintended purposes. The typical example is to link CET-4 scores to graduation and employment decision. When the assessment uses are beyond its intended purposes, the test is likely to be strongly criticized for its content or format. However, it should be noted that unintended consequences might result from flaws in test design or from other factors. For instance, a well-designed test may still generate negative effects because of misuses imposed on it. Therefore, test designers should not be solely blamed and bear full responsibility for detrimental consequences of the test. Test designers should be accountable for informing test takers and users of any negative influences they can anticipate even at the initial stage of test development.

The dimensions of assessment use are complex. Improving the test validity and ensuring qualities of test design are essential for appropriate use of an assessment. Only when test scores can be interpreted as valid, reliable, and meaningful indicator of test takers' language ability, can the assessment-based interpretations provide information that is meaningful, impartial, generalizable, relevant and sufficient to make the required decisions. Actually more issues at the decision-making level are involved, which may lead to unintended uses and negative consequences. Therefore, once a decision is made, decision makers and test users should be accountable for stakeholders by providing backing to justify its use. Factors involved in decision-making and the corresponding consequences deserve investigation. The investigation should take into account the societal and cultural values, university regulations and the common community practices. Test developers and decision-makers should work together to improve the assessment use. In other words, if universities insist on the policy of linking CET-4 scores to graduation decision, they are responsible for justifying this decision with evidence. The fundamental reason is that setting CET-4 scores as the threshold for awarding students their Bachelors' degree is not what the test is designed for. For another, it is ethical for test users to justify their decisions and to be held accountable to any consequences of their decisions.

One of the implications regarding washback phenomenon is also identified in this study. Teachers generally held positive and favorable attitudes to the current CET-4. They had a good understanding of the rationale and practical reasons for the 2006 CET reform. They also indicated their willingness to make changes in their teaching in line with the teaching objectives stipulated in both revised teaching and testing syllabuses. However, in comparison of teachers' responses with students' report about classroom teaching activities, it was found that teaching methods and contents seemed to remain unchanged. This finding confirmed the conclusions reached by previous studies involving teacher variables in washback mechanism (Cheng, 2005; Qi, 2005). For one thing, it may be because the CET-4 failed to initiate teachers' motivations to change their long-established teaching patterns given the short period of its administration. For another, washback mechanism is more complicated than what we assumed. Factors interweave together either facilitating or impeding washback phenomenon. It has to be cautioned that expectations on initiating pedagogical reforms by introducing test reform seemed too idealistic. Its effect and success cannot be fully ensured since teacher factors underling the washback phenomenon are more related to teachers' pedagogical theories, teaching experience and their beliefs on testing and learning (e.g., Alderson & Wall, 1993; Andrews, 2004; Cheng, 2005; Huang, 2004; Qi, 2005; Wall, 2005).

A positive indication is from teachers' evaluations of the effective strategies students used in their test taking process to improve test scores. Take listening for instance, teachers encouraged the strategies of "taking notes to help memory load" and "catching important conjunctions words to infer speakers' attitudes". The results from statistic analyses confirmed these two strategies as predicators of students' CET-4 scores. The congruence between what teachers intend to teach and what test developers intend to measure served as a foundation of positive washback of the CET-4. Only when teachers have a better understanding of the intended test purposes and test construct, there are chances that positive consequences of the test on teaching and learning can be maximized.

10.4 Limitations of the study

Although this study has used an AUA, a new framework to link validity and test consequence in a logical and coherent order, it should be admitted that it has some limitations.

The study only obtained CET-4 total scores and profile scores of the participants rather than subsection scores of different task types under each component. Thus, advanced statistical procedures like structural equation model cannot be employed to reveal the relationship between test components and overall English proficiency to be assessed, as well as the relationship among the four underlying traits (listening, reading, writing and translation abilities). This constitutes a major limitation of the present study.

Even though the questionnaires helped obtain information from large samples and the interviews helped provide in-depth findings, both generated self-reported data. Divergences may be identified between what respondents report and what they actually do. For instance, teachers reported that they would make some changes in teaching in accordance with the revised elements in the CET-4. However, what changes have been made and to what extent they adjusted their teaching were unknown. Thus, classroom observations could be conducted for qualitative description of real classroom activities. Due to limited time and resources, the study only explored classroom teaching and learning activities from students' and teachers' self-reported data. Lack of classroom observation is another limitation in this study.

The study based its investigation into the construct of the CET-4 on constructs defined in the CECR and the 2006 CET-4 syllabus. In test content analyses, language abilities and strategies intended to be examined by listening and reading tasks were checked in alignment with those listed in teaching and testing syllabuses. However, the study did not make further comparisons between skills coverage found in test content analyses with skills and strategies demonstrated by test takers in real test-taking processes. In this sense, the study may not present a larger picture or broad findings about validity issues.

Finally, the study was mainly conducted with participants from four sampled universities in Xi'an. Findings and conclusions can be representative of College English teaching and learning situations in university communities in Xi'an. However, it should be cautioned to generalize conclusions to the whole nation, because decisions made at institutional levels may vary greatly in light of different teaching qualities, resources, and students' varied English levels.

10.5 Suggestions for future research

It has been mentioned in Chapter 1 that the present study is to offer an exemplary attempt to link test validity and test consequences by drawing on Bachman and Palmer's AUA framework in China's higher education assessment context. The study demonstrated how to adapt the generic template of an AUA to a local and specific test. Therefore, one of the future directions is to replicate the present study within different testing settings and cultural contexts. Researchers can articulate their specific AUA frameworks to justify either large-scale and high-stakes tests like the CET-4 or small-scale and low-stakes tests like classroom assessment.

Second, this study adopted a bottom-up approach in the AUA to evaluate and examine uses and consequences of the CET-4. A reverse order starting with articulation of a consequence claim can be followed to develop an assessment. In fact, this top-down approach is more recommended by Bachman and Palmer because test purpose has always been the primary concern of test designers, and test consequences should be taken into account at the initial stage of test development. Thus, future studies are advised to demonstrate how an AUA framework serves as a logical and coherent guidance in developing an assessment.

Third, longitudinal studies on washback of the CET-4 are encouraged. Upon data collection of the study, the current CET-4 has been administered for only three years. Some nature and facets of CET-4 washback might not fully emerge or be perceived. In addition, stakeholders' perceptions of the CET-4 and evaluations of its washback on teaching and learning may change over time. Thus, this study contributes its value as a baseline study for future washback studies.

Another recommendation for future directions is pertinent to specificity and depth of a study. Under the guidance of the comprehensive and full-scale AUA framework, the study offered a panoramic view of validity and washback of the CET-4. Some issues definitely deserve in-depth investigation. Take stakeholders' attitudes for test preparation for example, universities preferred to open test preparation courses to ensure the overall CET-4 passing rate. Both teachers and students assumed that test preparations were more effective to improve test scores than test takers' English proficiency. Assumptions called for empirical research. Future studies can be conducted to investigate questions like: to what extent score gains can be achieved from test preparation activities, what are the underlying factors to facilitate score improvement, and what are the underlying factors to hinder students' English proficiency improvement.

Future studies can also be extended to a broader domain. This study focused on washback of the CET-4 within the instructional setting, but it does not mean social consequence of the CET-4 is a less important aspect for investigation. Since students' CET-4 scores serve as one of the criteria in making employment decisions, future studies can investigate how potential employers make recruitment decisions based on CET-4 scores.

Appendices

Appendix A Student questionnaire

A1. Chinese version of student questionnaire

学生问卷

各位同学，你们好！我们正在进行一项关于大学四级考试的调查。希望你能根据自己真实的想法和实际情况选择。你对问卷的认真填写将是对我们科研项目的大力支持。对于你提供的信息，我们将恪守保密原则，问卷将仅在科研范围内做统计和分析使用。

第一部分：　个人基本情况，请选择恰当答案打“√”

1. 性别　（１）男　　　　（２）　女

2. 年龄　（１）１８岁以下　（２）１８　（３）１９　（４）２０　（５）２１岁或以上

3. 专业:____________________(请填写)　（１）文科　（２）理科　(请选择)

4. 四级分数 ______________　(请填写)　　学生证号____________________________(请填写)

第二部分：　请按 5 个等级来评估以下问题，在对应的数字上打“√”。

A1.请评估四级各部分题型**难易度**	非常简单	简单	适中	难	非常难
四级总体	1	2	3	4	5
听力总体	1	2	3	4	5
短对话	1	2	3	4	5
长对话	1	2	3	4	5
听力短文	1	2	3	4	5
复合式听力	1	2	3	4	5
阅读总体	1	2	3	4	5
快速阅读	1	2	3	4	5
选词填空	1	2	3	4	5
阅读短文	1	2	3	4	5
完形填空	1	2	3	4	5
汉译英	1	2	3	4	5
写作	1	2	3	4	5

A2.四级各部分题型在何种程度上客观真实地**考察了您的英语综合能力？**	不能 反映	较小 程度	不确 定	一般 程度	较大 程度
四级总体	1	2	3	4	5
听力总体	1	2	3	4	5
短对话	1	2	3	4	5

长对话	1	2	3	4	5
听力短文	1	2	3	4	5
复合式听力	1	2	3	4	5
阅读总体	1	2	3	4	5
快速阅读	1	2	3	4	5
选词填空	1	2	3	4	5
阅读短文	1	2	3	4	5
完形填空	1	2	3	4	5
汉译英	1	2	3	4	5
写作	1	2	3	4	5

A3.	我参加四级考试是为了	完全不同意	不同意	不确定	同意	完全同意
A3.1.	检查我的英语水平	1	2	3	4	5
A3.2.	修学分	1	2	3	4	5
A3.3.	获得学位证	1	2	3	4	5
A3.4.	获得就业优势	1	2	3	4	5

		完全不同意	不同意	不确定	同意	完全同意
A4	将学位证与四级成绩挂钩可以促进我的英语学习	1	2	3	4	5
A5	用四级成绩单取代证书，可以更准确反映我的英语水平	1	2	3	4	5
A6	四级考试总体来说题型和内容是理想的	1	2	3	4	5
A7	听力试题的设计与现实生活中真实听力情景相似	1	2	3	4	5
A8	如果能先看到问题再听文章，会有助于我的听力理解	1	2	3	4	5
A9	阅读试题的设计与现实生活中真实阅读情景相似	1	2	3	4	5
A10	阅读理解的文章如果加上标题，会对我抓住主题预测内容有所帮助	1	2	3	4	5
A11	四级阅读文章主题较广泛，涉及人物文化地理历史教育科普等	1	2	3	4	5
A12	四级阅读文章体裁较单一，主要以议论文、叙述文、说明文为主，**应该包括**信函、广告、说明书等应用文。	1	2	3	4	5
A13	四级口试应该向所有考生开放来检测学生的英语交流能力	1	2	3	4	5
A14	如果口试是必考项目,我会花更多时间和精力提高自己的口语能力	1	2	3	4	5

A15	在四级考试中，我的**听力理解**受到以下因素的影响：	完全不同意	不同意	不确定	同意	完全同意
A15.1	词汇量不够影响我的听力理解	1	2	3	4	5
A15.2	长句、难句影响我的听力理解	1	2	3	4	5
A15.3	不了解听力材料的背景知识影响我的听力理解	1	2	3	4	5
A15.4	我觉得听懂了文章但听到问题时又不确定答案	1	2	3	4	5
A15.5	听力技巧策略欠缺影响我的听力理解	1	2	3	4	5

A16	在四级考试中，我的**阅读理解**受到以下因素的影响：	完全不同意	不同意	不确定	同意	完全同意
A16.1	词汇量不够影响我的阅读理解	1	2	3	4	5
A16.2	长句、难句影响我的阅读理解	1	2	3	4	5
A16.3	不了解阅读材料的背景知识影响我的阅读理解	1	2	3	4	5
A16.4	阅读速度较慢影响我的成绩	1	2	3	4	5
A16.5	阅读技巧欠缺影响我的阅读理解	1	2	3	4	5
A16.6	我觉得读懂了文章但又不能确定答案，个别选项迷惑性很大	1	2	3	4	5

A17.	做四级**听力理解**时	从不	偶尔	有时	经常	总是
A17.1	我会先看选项预测要听到的内容	1	2	3	4	5
A17.2	我会边听边做笔记帮助记忆特定细节	1	2	3	4	5
A17.3	遇到到生词，我会跳过生词继续听下去	1	2	3	4	5
A17.4	我会注意说话人的重音和语音语调来推断其态度或意图	1	2	3	4	5
A17.5	我会注意句间表示原因目的转折等的连接词来判断说话人观点	1	2	3	4	5

A18.	做四级**阅读理解**时	从不	偶尔	有时	经常	总是
A18.1	我会先看问题，再有目的的寻找答案	1	2	3	4	5
A18.2	我会快速浏览全文，抓住大意后再仔细阅读	1	2	3	4	5
A18.3	当我不懂句子意思时，我会分析句子的语法结构	1	2	3	4	5
A18.4	遇到生词，我会通过词根构词法来猜测词义	1	2	3	4	5
A18.5	遇到生词，我会通过上下文来猜测词义	1	2	3	4	5
A18.6	我会跳过生词继续阅读	1	2	3	4	5
A18.7	略读时，我能注重段落中心句获取文章大意，忽略细节	1	2	3	4	5
A18.8	查读时，我能带着问题获取特定信息	1	2	3	4	5

A19.	以下**何种因素**会**影响**到你的**四级整体成绩?**	不会影响	较小程度	不确定	一般程度	较大程度
A19.1.	语言难度	1	2	3	4	5
A19.2.	问题较难	1	2	3	4	5
A19.3.	陌生的文章主题背景	1	2	3	4	5
A19.4.	考试技巧策略欠缺	1	2	3	4	5
A19.5.	考试时间内不能完试题	1	2	3	4	5
A19.6.	考试焦虑	1	2	3	4	5

A20.	在以下哪方面**你受到四级成绩的影响?**	不会影响	较小程度	不确定	一般程度	较大程度
A20.1.	学习兴趣	1	2	3	4	5
A20.2.	成就感	1	2	3	4	5

A20.3.	自我评估	1	2	3	4	5
A20.4.	自信心	1	2	3	4	5
A20.5.	在同学中的形象	1	2	3	4	5
A20.6.	在老师心中的印象	1	2	3	4	5

		不能	较小	不确	一般	较大
			程度	定	程度	程度
B1	备考四级能提高我的四级分数	1	2	3	4	5
B2	备考四级能提高我的英语综合能力	1	2	3	4	5

		完全不	不同	不确	同	完全
		同意	意	定	意	同意
B3	做四级真题和模拟试题可以让我熟悉四级考试题型和内容	1	2	3	4	5
B4	做四级真题和模拟试题可以帮我检查自己学习的强项和不足之处	1	2	3	4	5
B5	我认为有必要停止正常教学进度进行备考复习	1	2	3	4	5
B6	我在参加大学英语四级考试这学期是为考试而学习	1	2	3	4	5

B7.	你们的**四级辅导课**从事以下活动的频繁程度如何？	从不	偶尔	有时	经常	总是
B7.1	提供关于四级考试题型内容的信息	1	2	3	4	5
B7.2	讲解真题和模拟试题	1	2	3	4	5
B7.3	讲解考试策略和技巧	1	2	3	4	5
B7.4	对各项语言技能进行分项练习	1	2	3	4	5

B8.	在**备考四级期间**我**平均每周花在**以下复习项目的时间:	从不	少于	2-3	4-5	5 小时
			1 小时	小时	小时	以上
B8.1.	复习语法	1	2	3	4	5
B8.2.	背单词	1	2	3	4	5
B8.3.	做真题	1	2	3	4	5
B8.4.	做模拟试题	1	2	3	4	5
B8.5.	学习考试策略技巧	1	2	3	4	5
B8.6.	练习听力	1	2	3	4	5
B8.7.	练习阅读	1	2	3	4	5
B8.8.	练习写作	1	2	3	4	5
B8.9.	练习翻译	1	2	3	4	5
B810.	练习口语	1	2	3	4	5

		完全不	不同	不确	同	完全
C1.	我在大学**学习英语**的主要**动机**是：	同意	意	定	意	同意
C1.1.	为了提高英语水平	1	2	3	4	5
C1.2	为了修学分	1	2	3	4	5

C1.3.　为了通过四级考试	1	2	3	4	5
C1.4.　为未来学习深造打基础	1	2	3	4	5
C1.5.　为了获得就业优势	1	2	3	4	5
C1.6.　为了满足社会需求	1	2	3	4	5

C2.　我在**课外**通过以下活动来提高我的英语能力	从不	偶尔	有时	经常	总是
C2.1　收听英文广播节目	1	2	3	4	5
C2.2　收看英文电视电影	1	2	3	4	5
C2.3　做四级听力理解来提高我的听力能力	1	2	3	4	5
C2.4　阅读英文报纸、杂志或小说	1	2	3	4	5
C2.5　做四级阅读理解来提高我的阅读能力	1	2	3	4	5
C2.6　用英语记笔记，写信或日记	1	2	3	4	5
C2.7　参加英语角或英语沙龙	1	2	3	4	5

C3.　你的老师在**英语课**上从事以下活动的频率：	从不	偶尔	有时	经常	总是
C3.1　讲解语言点(单词、句型)	1	2	3	4	5
C3.2　讲解课文	1	2	3	4	5
C3.3　讲解课后习题	1	2	3	4	5
C3.4　讲解和四级相关的内容	1	2	3	4	5
C3.5　讲解学习技巧和考试策略	1	2	3	4	5
C3.6　组织小组讨论	1	2	3	4	5
C3.7　组织语言游戏	1	2	3	4	5
C3.8　组织全班的综合语言活动	1	2	3	4	5

C4.　大学英语教学有待改善的是:	完全不	不同	不确	同	完全
	同意	意	定	意	同意
C4.1　提高大学生当前的英语交流能力，特别是听说能力	1	2	3	4	5
C4.2　班级人数过多	1	2	3	4	5
C4.3　教材和教学资料不够丰富	1	2	3	4	5
C4.3　现代化教学手段和设备不足	1	2	3	4	5
C4.5　大学英语教学课时不足	1	2	3	4	5

第三部分　单项选择　请在认为合适的选项序号上打“√”。

A21. 你所在的学校是否将学位证与四级成绩挂钩

(1). 是　　(2).不是　　(3).不确定

B9. 你自己通常何时开始进行备考复习

(1).第一学期　(2). 第二学期　(3). 第三学期　(4). 第四学期　(5).不复习

B10. 你大约做过多少套四级全真和模拟试题

(1).1-5 份　　(2). 6-10 份　　(3).11-15 份　　(4).16-20 份　　(5).21 份或以上

B11. 你的英语老师在**四级辅导课上**主要用何种语言

(1).只用英语　　(2).大部分用英语偶尔用中文解释　　(3).中英文对半　　(4).主要用中文

C5. 除了上英语课和备考外，你**每周平均**用于**课外英语**兴趣学习的时间

(1).少于 1 小时　(2).1-2 小时　　(3). 3-4 小时　(4).5-6 小时　　(5).7 小时以上

第四部分　排序题，　请将数字 1-5 填写在相对应的选项下

B12.请将备考四级的常用方法按重要性或有效性**排序**：**最重要（1）——最不重要（5）**

复习课本　　背单词　　题海战术　　学习应试技巧　　上辅导班

（ ）　　（ ）　　（ ）　　（ ）　　（ ）

C6.你认为英语 5 项技能按**重要性依次是:（1= 最重要 —— 5= 最不重要）**

听　　说　　读　　写　　译

（ ）　　（ ）　　（ ）　　（ ）　　（ ）

C7.你的强项依次是：**（1=最强 —— 5= 最弱）**

听　　说　　读　　写　　译

（ ）　　（ ）　　（ ）　　（ ）　　（ ）

衷心感谢您的耐心配合，请再检查一次是否每道题都做了。

谢谢！

A2. English version of student questionnaire

Dear students,

We are conducting a study on the CET-4. We would be appreciative if you could complete the questionnaire according to what your own opinion and **what you actually do** but not what you think should be done. There are no 'right' or 'wrong' answers. Your responses to this questionnaire will be treated in the strictest confidence and used only for the research purpose. Thank you very much for your cooperation.

Part One Please tick the appropriate answer

1.Your gender: (1) Male (2) Female

2. Your age: (1) below 18 (2) 18 (3) 19 (4) 20 (5) 21 or above

3. Your major: _______________(please specify) (1) Humanities & Arts (2) Science & Engineering

4. Your CET-4 scores _________ (please specify) or You Student ID _______________(please specify)

Part Two: Please grade the following statements on a 5-point scale.

A1. Please grade the **difficulty level** of different components of the CET-4 on a 5-polint Likert scale of difficulty: 1= very easy 2= easy 3= average 4= difficult 5= very difficult

The overall CET-4	1	2	3	4	5
The overall listening subtest	1	2	3	4	5
Short conversations	1	2	3	4	5
Long conversations	1	2	3	4	5
Listening passages	1	2	3	4	5
Compound dictation	1	2	3	4	5
The overall reading subtest	1	2	3	4	5
Fast reading	1	2	3	4	5
Banked cloze	1	2	3	4	5
In-depth reading passages	1	2	3	4	5
Cloze	1	2	3	4	5
Translation from Chinese to English	1	2	3	4	5
Essay writing	1	2	3	4	5

A2. Please evaluate the extent to which the CET-4 serves as **a measure of students' overall English proficiency**: 1= not at all 2= slightly 3= uncertain 4= to some extent 5= to a large extent

The overall CET-4	1	2	3	4	5
The overall listening subtest	1	2	3	4	5

Short conversations	1	2	3	4	5
Long conversations	1	2	3	4	5
Listening passages	1	2	3	4	5
Compound dictation	1	2	3	4	5
The overall reading subtest	1	2	3	4	5
Fast reading	1	2	3	4	5
Banked cloze	1	2	3	4	5
In-depth reading passages	1	2	3	4	5
Cloze	1	2	3	4	5
Translation from Chinese to English	1	2	3	4	5
Essay writing	1	2	3	4	5

Please evaluate the following items (A3-A16) based on a 5-point Likert scale of agreement:

1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree

A3	**The major aims for my taking the CET-4:**	SD	D	U	A	SA
A3.1	To check my English proficiency	1	2	3	4	5
A3.2	To satisfy academic credit requirement	1	2	3	4	5
A3.3	To get the Bachelor's degree	1	2	3	4	5
A3.4	To obtain advantage in employment	1	2	3	4	5
		SD	D	U	A	SA
A4	Setting the CET-4 cut-off score as a prerequisite for a Bachelor's degree can push me to learn English	1	2	3	4	5
A5	Compared with the CET-4 certificate, the new score report can reflect my English proficiency more accurately.	1	2	3	4	5
A6	The overall formats and contents of the CET-4 are satisfactory	1	2	3	4	5
A7	Listening tasks are similar to those in the real-life situations.	1	2	3	4	5
A8	Printing listening questions on paper can facilitate my comprehension.					
A9	Reading tasks are similar to those in the real-life situations.	1	2	3	4	5
A10	Adding titles to passages can facilitate my prediction of the content.	1	2	3	4	5
A11	Reading passages cover a wide range of topics such as humanities, culture, history, education, geography, and science, etc.	1	2	3	4	5
A12	Reading genres should be more diverse with practical passages such as letters, advertisements, and instructions besides the dominant genres of argumentation, narration and exposition.	1	2	3	4	5
A13	The CET-SET4 should be open to all test takers to check their ability to communicate in English orally.	1	2	3	4	5

A14	If the CET-SET4 is compulsory, I will spend more time and efforts cultivating my speaking ability.	1	2	3	4	5

A15.	My CET-4 **listening performance** is affected by	SD	D	U	A	SA
A15.1	Limited vocabulary	1	2	3	4	5
A15.2	Long and complex sentences	1	2	3	4	5
A15.3	Lack of background knowledge	1	2	3	4	5
A15.4	Confusion by the tricky options in spite of my understanding	1	2	3	4	5
A15.5	Lack of listening skills and strategies.	1	2	3	4	5

A16.	My CET-4 **reading comprehension** is affected by	SD	D	U	A	SA
A16.1	Limited vocabulary	1	2	3	4	5
A16.2	Long and complex sentences	1	2	3	4	5
A16.3	Lack of background knowledge	1	2	3	4	5
A16.4	Slow reading speed	1	2	3	4	5
A16.5	Lack of reading skills and strategies	1	2	3	4	5
A16.6	Confusion by the tricky options in spite of my understanding	1	2	3	4	5

Please evaluate A17 andA18 based on a 5-point Likert scale of frequency:

1= never, 2= seldom, 3= sometimes, 4= often, 5= always

A17.	When I take the CET-4 **listening subtest**,					
A17.1	I read options first to predict what I'm going to hear.	1	2	3	4	5
A17.2	I take notes to help my memory.	1	2	3	4	5
A17.3	I just skip unknown words so as to concentrate on the whole.	1	2	3	4	5
A17.4	I pay attention to the speakers' stress and intonation to infer their intentions or attitudes.	1	2	3	4	5
A17.5	I pay attention to conjunctions such as "but, so that…" to infer speakers' opinions.	1	2	3	4	5

A18.	When I take the CET-4 **reading subtest**,					
A18.1	I read questions first before reading passages.	1	2	3	4	5
A18.2	I look through the passage first for the main idea before my careful reading.	1	2	3	4	5
A18.3	I analyze grammatical structures to help me understand difficult and complex sentences.	1	2	3	4	5
A18.4	I guess the meaning of an unknown word by its root, prefix or suffix.	1	2	3	4	5

A18.5	I guess the meaning of unknown words from the context.	1	2	3	4	5
A18.6	I just skip unknown words and continue to focus on my reading.	1	2	3	4	5
A18.7	I skim to identify the main idea.	1	2	3	4	5
A18.8	I scan to search for the specific details.	1	2	3	4	5

Please evaluate A19 and A20 based on a 5-point Likert scale of frequency:

1= not at all, 2= slightly, 3= uncertain, 4= to some extent, 5= to a large extent

A19. My overall CET-4 test performance is affected by factors as follows:

A19.1.	Difficulty of language	1	2	3	4	5
A19.2.	Difficulty of questions	1	2	3	4	5
A19.3.	Unfamiliarity with topics	1	2	3	4	5
A19.4.	Lack of test-taking strategies	1	2	3	4	5
A19.5.	Time pressure	1	2	3	4	5
A19.6.	Test anxiety	1	2	3	4	5

A20. The CET-4 affects me in the following aspects:

A20.1.	Study interest	1	2	3	4	5
A20.2.	Sense of achievement	1	2	3	4	5
A20.3.	Self-evaluation	1	2	3	4	5
A20.4.	Self-confidence	1	2	3	4	5
A20.5.	Image among my classmates	1	2	3	4	5
A20.6.	Image among my teachers	1	2	3	4	5

B1	Test preparations can improve my CET-4 scores.	1	2	3	4	5
B2	Test preparations can improve my overall English proficiency.	1	2	3	4	5

		SD	D	U	A	SA
B3	Doing past or mock test papers can familiarize me with the CET-4 content and format.	1	2	3	4	5
B4	Doing past or mock test papers helps me diagnose my learning strengths and weaknesses.	1	2	3	4	5
B5	It’s necessary to stop the normal teaching for test preparation.	1	2	3	4	5
B6	I learn to the test during the semester in which I take CET-4.	1	2	3	4	5

B7. Please evaluate to what extent your CET-4 preparation course covers the following activities?

1= never, 2= seldom, 3= sometimes, 4= often, 5= always

B7.1	Offering information about test contents and the formats	1	2	3	4	5
B7.2	Explaining past or mock test papers	1	2	3	4	5
B7.3	Developing students' test taking strategies	1	2	3	4	5
B7.4	Doing exercises related to the five language skills	1	2	3	4	5

B8. Please evaluate the average amount of test preparation time you spend on the following per week:

1= never, 2= a little (less than 1 hour), 3= moderate (2-3 hours), 4= a lot (4-5 hours),
5=extensive (more than 5 hours)

B8.1.	Reviewing grammar	1	2	3	4	5
B8.2.	Memorizing vocabulary	1	2	3	4	5
B8.3.	Reviewing past test papers	1	2	3	4	5
B8.4.	Doing mock tests	1	2	3	4	5
B8.5.	Learning test taking strategies	1	2	3	4	5
B8.6.	Practicing listening	1	2	3	4	5
B8.7.	Practicing reading	1	2	3	4	5
B8.8.	Practicing writing	1	2	3	4	5
B8.9.	Practicing translation	1	2	3	4	5
B810.	Practicing speaking	1	2	3	4	5

C1	The major aims for my English learning at university:	SD	D	U	A	SA
C1.1	To further improve my English level	1	2	3	4	5
C1.2	To satisfy academic credit requirements	1	2	3	4	5
C1.3	To pass the CET-4	1	2	3	4	5
C1.4	To pursue further studies	1	2	3	4	5
C1.5	To obtain advantage in employment	1	2	3	4	5
C1.6	To satisfy social needs for graduates with higher English proficiency	1	2	3	4	5

C2 Outside of class I try to improve my English by:

1= never, 2= seldom, 3= sometimes, 4= often, 5= always

C2.1	Listening to radio programs in English	1	2	3	4	5
C2.2	Watching films and TV programs in English	1	2	3	4	5
C2.3	Doing CET-4 listening test to improve my listening ability	1	2	3	4	5
C2.4	Reading English newspapers, magazines and books.	1	2	3	4	5
C2.5	Doing CET-4 reading test to improve my reading ability	1	2	3	4	5

C2.6	Writing mails and diaries in English	1	2	3	4	5
C2.7	Practicing oral English at English corners or saloons	1	2	3	4	5

C3. How often does your teacher engage in the following classroom activities:

1= never, 2= seldom, 3= sometimes, 4= often, 5= always

C3.1	Explaining language points such as vocabulary and sentence structures	1	2	3	4	5
C3.2	Explain the meaning of the text	1	2	3	4	5
C3.3	Explaining textbook exercises	1	2	3	4	5
C3.4	Providing information or explaining test content related to the CET-4	1	2	3	4	5
C3.5	Explaining learning skills and test-taking strategies	1	2	3	4	5
C3.6	Organizing classroom activities such as pair work, group discussions	1	2	3	4	5
C3.7	Organizing language games	1	2	3	4	5
C3.8	Organizing integrated language activities	1	2	3	4	5

C4. The College English teaching needs improvement in the following aspects:

1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree.

C4.1	Students' ability to communicate, esp. their listening ability	1	2	3	4	5
C4.2	Large class size	1	2	3	4	5
C4.3	Inadequate textbooks and other available teaching resources	1	2	3	4	5
C4.4	The lack of teaching and learning aids and facilities	1	2	3	4	5
C4.5	Inadequate class hours per week	1	2	3	4	5

Part Three Please tick only ONE appropriate answer

A21. Does your university set the CET-4 cut-off score as a prerequisite of a Bachelor's degree?

(1). Yes (2). No (3). uncertain

B9. Your preparation for the CET-4 usually begins from:

(1). 1st semester (2). 2nd semester (3). 3rd semester (4). 4th semester (5). No reparation

B10. How many past and mock test papers have you done for the test preparation?

(1). 1-5 (2). 6-1 (3). 11-15 (4). 16-15. (5). 15. or above

B11. What is the medium of instruction your teacher uses in the *test preparation course*?

(1). English only (2). English with occasional Chinese

(3). Half English half Chinese (4). Mainly Chinese

C5. In addition to course study and test preparation, the average time you spent per week on the extra-curriculum English interest study is____.

(1). less than 1 hour (2). 1-2 hours (3). 3-4 hours (4). 5-6 hours (5). more than 7 hours

Part Four Rank ordering

Please write down the number according to their ranks with *1 as the most or the best and 5 as the least or the worst* based on your personal judgment.

B12. Please put the following test preparation methods in an order (1-5) with *1 being the most important and effective.*

To review textbooks	To memorize vocabulary	To do exercises	To learn test taking strategies	To attend tutorial course
（　）	（　）	（　）	（　）	（　）

C6. Please put the five language skills that are most important for a college student to cultivate in an order (1-5) with 1 being the most important.

Listening	Speaking	Reading	Writing	Translation
（　）	（　）	（　）	（　）	（　）

C7. Please put the five language skills in an order (1-5) to reflect your strength, with 1 being your strongest skill

Listening	Speaking	Reading	Writing	Translation
（　）	（　）	（　）	（　）	（　）

End of the questionnaire

Thanks your cooperation!

Appendix B Teacher Questionnaire

Dear colleagues,

We are conducting a study on college English teaching and learning and the CET-4. We would be appreciative if you could complete the questionnaire. Your responses to this questionnaire will be treated in the strictest confidence and used only for the research purpose. Thank you very much for your cooperation.

Part One Please tick the appropriate answer

1.Your gender: (1) Male (2) Female

2. Your age: (1) below 25 (2) 26-35 (3) 36-45 (4) 46-55 (5) 56 or above

3. Your academic qualifications: (1) Below BA (2) BA (3) MA (4) PhD

4. Your professional qualifications:

(1) teaching assistant (2)lecturer (3) associate professor (4) Professor (5) Professor& doctoral supervisor

5. Years of your tertiary teaching experience

(1) less than 5 years (2) 6-10 (3) 11-20 (4) 21-30 (5) 31 years or above

6. Students you are currently teaching: (1) Freshman (2) Sophomore

7. Number of hours you teach per week (only your primary workload):

(1) less than 8 hours (2) 9-10 (3) 11-12 (4) 13-14 (5) 15 or above

Part Two Please grade the following sayings on a 5-point Likert scale.

A1. The major reasons you perceive behind the CET-4 reform in 2006:

1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree

A1.1. To meet the demands of tertiary education 1 2 3 4 5

A1.2. To positively impact the college English teaching and learning 1 2 3 4 5

A1.3. To refine testing methods 1 2 3 4 5

A1.4. To further improve the CET-4 as a measure of students' English proficiency 1 2 3 4 5

A1.5. To motivate students to lay more emphasis on listening ability 1 2 3 4 5

A1.6. To meet social needs for graduates with higher English proficiency 1 2 3 4 5

A1.7. To prepare students for their future career 1 2 3 4 5

A2. Please evaluate the major changes you have perceived in the new CET-4 on a 5-point Likert scale of emphasis: 1= A lot less emphasis, 2= Somewhat less emphasis, 3= No change, 4= somewhat more emphasis, 5= a lot more emphasis

A2.1. Emphasis on listening 1 2 3 4 5

A2.2. Emphasis on integrated skills 1 2 3 4 5

A2.3. Emphasis on authenticity 1 2 3 4 5

A2.4. Emphasis on being communicatively-oriented	1	2	3	4	5
A2.5. Emphasis on productive skills	1	2	3	4	5
A2.6. Emphasis on reading	1	2	3	4	5
A2.7. Emphasis on grammatical usage	1	2	3	4	5

Please evaluate item A4 and A4 based on a 5-point Likert scale of agreement:

1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree

A3. The major changes you are likely to make in your teaching in the context of the new CET-4	SD	D	U	A	SA
A3.1. To teach in accordance with the new test formats & contents	1	2	3	4	5
A3.2. To adopt new teaching methods	1	2	3	4	5
A3.3. To use a more communicative teaching approach	1	2	3	4	5
A3.4. To lay more emphasis on developing students' listening ability	1	2	3	4	5
A3.5. To lay more emphasis on developing students' fast reading ability	1	2	3	4	5
A3.6. To lay more emphasis on developing students' careful reading ability	1	2	3	4	5

A4. CET-4 has exerted a great effect on the following aspects in the past two decades:	SD	D	U	A	SA
A4.1. Improving students' linguistic competence	1	2	3	4	5
A4.2. Improving students' communicative competence	1	2	3	4	5
A4.3. Promoting college English teaching on the whole	1	2	3	4	5
A4.4. Inducing the phenomenon of "high marks, low abilities"	1	2	3	4	5
A4.5. Inducing the phenomenon of "teaching and learning to the test"	1	2	3	4	5

A5. Students' CET-4 scores or passing rates tend to affect you in the following aspects:
1= not at all, 2= slightly, 3= uncertain, 4= to some extent , 5= to a large extent

A5.1. Academic promotion	1	2	3	4	5
A5.2. Sense of achievement	1	2	3	4	5
A5.3. Self-evaluation	1	2	3	4	5
A5.4. Cash bonus	1	2	3	4	5
A5.5. Popularity with students	1	2	3	4	5
A5.6. Image among colleagues	1	2	3	4	5

A6. Please grade the **difficulty level** of different components of the CET-4 on a 5-polint Likert scale of difficulty: 1= very easy 2= easy 3= average 4= difficult 5= very difficult

The overall CET-4	1	2	3	4	5
The overall listening subtest	1	2	3	4	5
Short conversations	1	2	3	4	5
Long conversations	1	2	3	4	5
Listening passages	1	2	3	4	5
Compound dictation	1	2	3	4	5
The overall reading subtest	1	2	3	4	5
Fast reading	1	2	3	4	5
Banked cloze	1	2	3	4	5
In-depth reading passages	1	2	3	4	5
Cloze	1	2	3	4	5
Translation from Chinese to English	1	2	3	4	5
Essay writing	1	2	3	4	5

A7. Please evaluate the extent to which the CET-4 serves as **a measure of students' overall English proficiency**: 1= not at all 2= slightly 3= uncertain 4= to some extent 5= to a large extent

The overall CET-4	1	2	3	4	5
The overall listening subtest	1	2	3	4	5
Short conversations	1	2	3	4	5
Long conversations	1	2	3	4	5
Listening passages	1	2	3	4	5
Compound dictation	1	2	3	4	5
The overall reading subtest	1	2	3	4	5
Fast reading	1	2	3	4	5
Banked cloze	1	2	3	4	5
In-depth reading passages	1	2	3	4	5
Cloze	1	2	3	4	5
Translation from Chinese to English	1	2	3	4	5
Essay writing	1	2	3	4	5

Please evaluate the following items (A8 -A17) based on a 5-point Likert scale of agreement:
1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree

A8. The overall formats and contents of the CET-4 are satisfactory.	1	2	3	4	5
A9. Listening tasks are similar to those in the real-life situations.	1	2	3	4	5
A10. Printing listening questions on paper can facilitate students' understanding	1	2	3	4	5

A11. Reading tasks are similar to those in the real-life situations. 1 2 3 4 5

A12. Adding titles to passages can facilitate students' prediction of the content. 1 2 3 4 5

A13. Reading passages cover a wide range of topics such as humanities, culture, history, education, geography, and science, etc. 1 2 3 4 5

A14. Reading genres should be diversified with more practical passages such as letters, advertisements, and instructions besides the dominant genres of argumentation, narration and exposition. 1 2 3 4 5

A15. CET Spoken English Test (SET) should be open to all test takers so as to check their ability to communicate in English orally. 1 2 3 4 5

A16. If CET-SET is open to all test takers, my students will spend more time and efforts cultivating their speaking ability. 1 2 3 4 5

A17. The CET-4 score report can reflect students' English proficiency more accurately than the traditional certificate. 1 2 3 4 5

Please evaluate B1 and B2 based on a 5-point Likert scale of extent:
1= not at all, 2= slightly, 3= uncertain, 4= to some extent , 5= to a large extent

B1. Test preparations can improve students' scores. 1 2 3 4 5

B2. Test preparations can improve students' overall English proficiency. 1 2 3 4 5

B3. To what extent your CET-4 preparation courses cover the following activities?
1= never, 2= seldom, 3= sometimes, 4= often, 5= always

B3.1. Offering information about test contents and the formats 1 2 3 4 5

B3.2. Explaining authentic or mock test papers 1 2 3 4 5

B3.3. Developing students' test taking strategies 1 2 3 4 5

B3.4. Doing exercises related to the five language skills 1 2 3 4 5

Please evaluate the following items (B4 -C5) based on a 5-point Likert scale of agreement:
1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree.

B4. What do you think are the main functions of doing mock and authentic tests papers?

B4.1. To familiar students with the test formats and contents 1 2 3 4 5

B4.2. To diagnose students' learning strengths and weaknesses 1 2 3 4 5

C1. The short-term objective of College English teaching is to help students obtain high scores in CET-4. 1 2 3 4 5

C2. The long-term objective of College English teaching is to develop students' ability to use English in an all-round way. 1 2 3 4 5

C3. The College English teaching needs improvement in the following aspects:

C3.1. Students’ ability to communicate, esp. their listening ability	1	2	3	4	5
C3.2. Large class size	1	2	3	4	5
C3.3. Inadequate textbooks and other available teaching resources	1	2	3	4	5
C3.4. The lack of modern teaching and learning aids and facilities	1	2	3	4	5
C3.5. Inadequate class hours per week	1	2	3	4	5

C4. The factors that most influence your teaching are:

C4.1. Teaching belief	1	2	3	4	5
C4.2. Teaching experience	1	2	3	4	5
C4.3. Teaching syllabus	1	2	3	4	5
C4.4. CET-4	1	2	3	4	5
C4.5. Textbooks	1	2	3	4	5
C4.6. Students’ expectations	1	2	3	4	5
C4.7. Past experience as a language learner	1	2	3	4	5
C4.8. University’s curriculum requirement	1	2	3	4	5

C5. In the semester when students sit for the CET-4

C5.1. Teaching should still focus on the textbooks.	1	2	3	4	5
C5.2. Normal teaching should be suspended and replaced with test preparation courses.	1	2	3	4	5
C5.3. I skip over certain sections in the textbook to squeeze time for test preparation	1	2	3	4	5
C5.4. I teach what will be tested because my students want me to do so.	1	2	3	4	5

C6. How often do you engage in the following classroom activities: 1= never, 2= seldom, 3= sometimes, 4= often, 5= always

C6.1 Explaining language points such as vocabulary and sentences	1	2	3	4	5
C6.2 Explaining the meaning of the text	1	2	3	4	5
C6.3 Explaining textbook exercises	1	2	3	4	5
C6.4 Providing information or explaining test content related to the CET-4	1	2	3	4	5
C6.5 Explaining learning skills and test taking strategies	1	2	3	4	5
C6.6 Organizing pair work or group discussions	1	2	3	4	5
C6.7 Organizing language games	1	2	3	4	5
C6.8 Organizing integrated language activities	1	2	3	4	5

Please evaluate the following items (C7 –D4) based on a 5-point Likert scale of agreement:
1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree.

C7. In your teaching, you encourage students to do the following to improve their listening performance:

C7.1. Predicting the listening content by looking through the options	1	2	3	4	5
C7.2. Taking notes while listening to help them remember details	1	2	3	4	5
C7.3. Skip the unknown words so as to concentrate on the whole	1	2	3	4	5
C7.4. Paying attention to the stress and intonation to infer speaker's intentions or attitudes	1	2	3	4	5
C7.5. Paying attention to some conjunctions to infer speakers' opinion	1	2	3	4	5

C8. In your teaching, you encourage students to do the following to improve their reading performance

C8.1. Reading questions first before reading passages	1	2	3	4	5
C8.2.Looking through the passage for the main idea before careful reading	1	2	3	4	5
C8.3.Analyzing the grammatical structure for the difficult and complex sentences	1	2	3	4	5
C8.4.Guessing unknown words in the context or by the word-building knowledge	1	2	3	4	5
C8.5.Using skimming and scanning skills for different purposes	1	2	3	4	5

D1. The major aims for students' learning English at university:

D1.1. To further improve their English proficiency	1	2	3	4	5
D1.2. To satisfy the academic credit requirement	1	2	3	4	5
D1.3. To pass the CET-4	1	2	3	4	5
D1.4. To pursue further studies	1	2	3	4	5
D1.5. To obtain advantage in employment	1	2	3	4	5
D1.6. To meet the social needs for graduates with higher English proficiency	1	2	3	4	5

D2. What do you think are the major reasons for students' taking the CET-4?

D2.1 To check their English proficiency	1	2	3	4	5
D2.2 To get the academic credits	1	2	3	4	5
D2.3 To obtain the Bachelor's degree	1	2	3	4	5
D2.4 To obtain advantage in employment	1	2	3	4	5

D3. Please evaluate the average amount of test preparation time your students spend on the following per week: 1= never, 2= a little (less than 1 hour), 3= moderate (2-3 hours),
4= a lot (4-5 hours), 5=extensive (more than 5 hours)

D3.1.Reviewing grammar	1	2	3	4	5
D3.2.Memorizing vocabulary	1	2	3	4	5

D3.3.Reviewing past test papers	1	2	3	4	5
D3.4.Doing mock tests	1	2	3	4	5
D3.5. Learning test taking strategies	1	2	3	4	5
D3.6. Practicing listening	1	2	3	4	5
D3.7. Practicing reading	1	2	3	4	5
D3.8. Practicing writing	1	2	3	4	5
D3.9. Practicing translation	1	2	3	4	5
D3.10. Practicing speaking	1	2	3	4	5

Please evaluate the following items (C7 –D4) based on a 5-point Likert scale of agreement:
1= strongly disagree, 2= disagree, 3= uncertain, 4= agree, 5= strongly agree.

D4. Students' CET-4 listening performance is affected by the following factors:

D4.1. Limited vocabulary.	1	2	3	4	5
D4.2. Long and complex sentences.	1	2	3	4	5
D4.3. Lack of background knowledge	1	2	3	4	5
D4.4. Lack of listening skills and strategies	1	2	3	4	5
D4.5. Confusion by the tricky options in spite of students' understanding	1	2	3	4	5

D5. Students' CET-4 reading performance is affected by the following factors:

D5.1. Limited vocabulary.	1	2	3	4	5
D5.2. Long and complex sentences.	1	2	3	4	5
D5.3. Lack of background knowledge	1	2	3	4	5
D5.4. Lack of reading skills and strategies	1	2	3	4	5
D5.5. Slow reading speed	1	2	3	4	5
D5.6. Confusion by the tricky options in spite of their understanding the passages	1	2	3	4	5

D6. My overall CET-4 test performance is affected by factors as follows: 1= not at all,
2= slightly, 3= uncertain, 4= to some extent, 5= to a large extent

D6.1 Difficulty of language	1	2	3	4	5
D6.2 Difficulty of questions	1	2	3	4	5
D6.3 Unfamiliarity with topics	1	2	3	4	5
D6.4 Lack of test-taking strategies	1	2	3	4	5
D6.5 Time pressure	1	2	3	4	5
D6.6 Test anxiety	1	2	3	4	5

Part Three   Please tick only ONE appropriate answer

B5. In your university students can take the CET-4 in

(1) 1st semester (2) 2nd semester (3) 3rd semester (4) 4th semester

B6. The preparation for CET-4 in your school usually begins from:

(1) 1st semester (2) 2nd semester (3) 3rd semester (4) 4th semester

B7. How many authentic and mock test papers do you usually review in the test preparation course?

(1) 0-5 (2) 5-10 (3) 11-15 (4) 16-20 (5) 21or above

B8. What is the medium of instruction you use in your test preparation course?

(1) English only (2) English with occasional Chinese

(3) Half English half Chinese (4) Mainly Chinese

C9. The typical size of your classes in terms of student numbers

(1) 20-30 (2) 31-50 (3) 51-70 (4) 71-90 (5) 90 or above

C10. Are you satisfied with the textbooks you are currently using?

(1) very dissatisfied (2) dissatisfied (3) uncertain (4) satisfied (5) very satisfied

End of the questionnaire

Thanks for your cooperation!

Appendix C Needs analysis in preliminary study

C1. Chinese version of needs analysis

我在日常学习和未来的工作生活中可能会遇到要求使用英语的场合		从不	偶尔	有时	经常	总是
1	上课，听讲座	1	2	3	4	5
2	英语角	1	2	3	4	5
3	参加考试	1	2	3	4	5
4	工作面试	1	2	3	4	5
5	学术交流	1	2	3	4	5
6	获取用英语发布的信息（新闻，网站）	1	2	3	4	5
7	看英文电视电影	1	2	3	4	5
8	出国旅游	1	2	3	4	5
9	与外国人交流	1	2	3	4	5

我在日常学习和未来的工作生活中可能阅读到英文		从不	偶尔	有时	经常	总是
10	教科书	1	2	3	4	5
11	考试题	1	2	3	4	5
12	文献论文	1	2	3	4	5
13	专业书籍	1	2	3	4	5
14	报纸杂志	1	2	3	4	5
15	小说	1	2	3	4	5
16	说明书	1	2	3	4	5
17	合同	1	2	3	4	5

我在日常学习和未来的工作生活中可能会用英文写		从不	偶尔	有时	经常	总是
18	日记	1	2	3	4	5
19	论文摘要	1	2	3	4	5
20	电子邮件	1	2	3	4	5
21	会议备忘录	1	2	3	4	5

C2. English version of needs analysis

Please evaluate the following items based on a 5-point Likert Scale of frequency:
1= never, 2= occasionally, 3= sometimes, 4= often, 5= always.

The circumstances in which I may use English in daily or future work-related interactions

1	To attend lectures delivered in English	1	2	3	4	5
2	To go to English corner	1	2	3	4	5
3	To take test	1	2	3	4	5
4	To attend job interviews	1	2	3	4	5
5	To engage in academic exchanges	1	2	3	4	5
6	To acquire information from news reports and websites in English	1	2	3	4	5
7	To watch TV programs and films	1	2	3	4	5
8	To travel abroad	1	2	3	4	5
9	To communicate with foreigners	1	2	3	4	5

The text types I may read in English in daily or future work-related interactions

10	Text books	1	2	3	4	5
11	Test exercises	1	2	3	4	5
12	Literature review	1	2	3	4	5
13	Professional books	1	2	3	4	5
14	Newspapers and magazines	1	2	3	4	5
15	Novels	1	2	3	4	5
16	Instructions	1	2	3	4	5
17	Contracts	1	2	3	4	5

The text types I may write in English in daily or future work-related interactions

18	Diary	1	2	3	4	5
19	Thesis abstract	1	2	3	4	5
20	Email	1	2	3	4	5
21	memo	1	2	3	4	5

References

Alderson, J.C. (1991). Dis-sporting life. Response to Alistair Pollitt's paper 'Giving students a sporting chance.' In J. Alderson & B. North (Eds.), *Language Testing in the 1990s*. London: Macmillan Publishers.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J.C., Clapham, C.M., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J.C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3): 280-297.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*(2), 253-270.

Alderson, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics 14*(2): 115-129.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement*, *51*(2) 1-38.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and Manuals.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests and Manuals*. Washington, DC: American Psychological Association.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Reviews of Psychology*, *37*(1), 1-16.

Anastas, J. W. (1999). *Research design for social work and the human services* (2nd ed.). New York: Columbia University Press.

Andrews, S. (1994). The washback effect of examinations: Its impact upon curriculum innovation in English language teaching. *Curriculum Forum, 4*(1), 44-58.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37–52). Mahwah, NJ: Lawrence Erlbaum.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.9-13). Hillsdale, NJ: Lawrence Erlbaum.

Babbie, E.R. (2007). *The practice of social research* (11th ed.). Belmont: Wadsworth Publishing Company.

Bingham, W.V. (1937). *Aptitudes and aptitude testing*. New York: Harper.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*(3), 5-18.

Bachman, L. F. (2003). Constructing an assessment use argument and supporting claims about test taker-assessment task interactions in evidence-centered assessment design. *Measurement: interdisciplinary Research and perceptive*, 1, 63-65.

Bachman, L. F. (2004a). Linking observations to interpretations and uses in TESOL research. *TESOL Quarterly, 38*(4), 723-728.

Bachman, L. F. (2004b). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F. (2007). Language assessment: Opportunities and challenges. *Meeting of the American Association of Applied Linguistics (AAAL), Costa Mesa, CA*.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language* (Vol. 1). Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1982). The Construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449-465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in the Real World: Developing Language assessments and justifying their use*. Oxford: Oxford University Press.

Bachman, L. F., Vanniaraian, A. K. S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5*(2), 128-159.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Bentler, P. M. (1995). *EQS structural equations program manual* (revised ed). Los Angeles, CA: BMDP Statistical Software, Inc.

Brennan. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger.

Brewer, J., & Hunter, A. (1989). *Multimethod research: A synthesis of styles*. Sage Publications, Inc.

Brown, G. T., & Hirschfeld, G. H. (2008). Students’ conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice*, *15*(1), 3-17.

Brown, J.D. (1996). *Testing in language programs*. Upper saddle River, NJ; Prentice Hall.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Bryman, A., & Cramer, D. (2012). *Quantitative data analysis with IBM SPSS 17, 18 & 19: A guide for social scientists.* London: Routledge.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8*(1), 67-91.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Burns, R. B. (2000). *Introduction to research methods* (4th ed.). Pearson Education Australia: Longman.

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contents and methods* (pp. 113-128). Mahwah, NJ: Lawrence Erlbaum.

Cai, J. (2002). Influence of CET writing requirements and scoring criteria on Chinese students’ composition. *Journal of PLA University of Foreign Languages*, *25*(5), 49-53.

Cai, J. (2005). Some thoughts on College English teaching. *Foreign Language Teaching and Education*, *37*(2), 83-91.

Cai, J. (2006). *EFL at tertiary level in China: Review, reflection and research*. Shanghai, China: Fudan University press.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. Oller (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, Mass.: Newbury House.

College English Test (2011). What is the test about? Retrieved Nov. 16 from www.en.cet.edu.cn/

Chapelle, C.A. (1999). Validation in language assessment. *Annual Review of Applied Linguistics*, *19*(1), 254-272.

Chapelle, C.A., Enright, M. K. & Jamieson, J. (2004). Issues in developing a TOEFL validity argument. Paper presented at the language Testing Research Colloquium, Temecula, CA.

Chapelle, C.A., Enright, M.K., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chen, J. F., Warden, C. A., & Chang, H. T. (2005). Motivators that do not motivate: The case of Chinese EFL learners and the influence of culture on motivation. *TESOL Quarterly, 39*(4), 609-633.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38-54.

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation, 24*(3), 279-301.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Contents and Methods* (pp. 147-170). Mahwah, NJ: Lawrence Erlbaum.

Cheng, L. (2005). *Changing language teaching through language testing*. Cambridge: Cambridge University Press.

Cheng, L. (2008). The key to success: English language testing in China, *Language Testing, 25*(11), 15-37.

Cheng, L. (2010). The history of examinations why, how, what and whom to select? In L. Cheng & A. Curtis (Eds.), *English Language Assessment and the Chinese Learner* (pp. 44-59). New York, NY: Routledge.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contents and methods* (pp.3-17). Mahwah, NJ: Lawrence Erlbaum.

Cheng, L., & Curtis, A. (2010). *English language assessment and the Chinese learner*. New York, NY: Routledge.

Cheng, L., & Qi, L. (2006). Description and examination of the national matriculation English test. *Language Assessment Quarterly: An International Journal, 3*(1), 53-70.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods.* Mahwah, NJ: Lawrence Erlbaum.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension* (Vol. 4), Cambridge: Cambridge University Press.

Cohen, A. D. (1980). *Testing language ability in the classroom*. Rowley, MA: Newbury House Publishers.

Cohen, A.D. (1984). On taking tests: What the students report. *Language Testing, 1*(1): 70-81.

Cohen, A. (1998). *Strategies in learning and using a second language*. London: Longman.

Cohen, A., & Macaro, E. (2008). *Language learner strategies: Thirty years of research and practice*. Oxford: Oxford University Press.

Cohen, A., & Upton, T. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph No. MS-33). Princeton, NJ: Education Testing Service.

Cohen, L. & Manion, L. (1991). *Research methods in education* (3rd ed.). London: Routledge.

Cohen, L.,& Manion, L., & Morrison, K.(2000). *Research methods in education* (5th ed.). London: Routledge.

Corson, D. (1997). *Encyclopedia of language and education* (Vol. 8). Kluwer Academic Publishers.

Cronbach, L.J. (1949). *Essentials of Psychological testing*. New York: Harper.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J. (1980a). Selection theory for a political world. *Public Personnel Management*, *9*(1), 37-50.

Cronbach, L.J. (1980b). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement over a Decade*, 5, 99-108.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L.J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Cumming, A. (1996). Introduction: The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing*. Clevedon, UK: Multilingual Matters.

Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.

Deci, E. L., Connell, J. P., & Ryan, R. M. (1989). Self-determination in a work organization. *Journal of Applied Psychology*, 74, 580-590.

Denzin, N. K. (1970). *The research act*. Chicago: Aldine.

Denzin, N.K. (1994). The art and politics of interpretation. In N. K. Denzin & Y.S. Lincoln (Eds.), *Handbook of qualitative research* (pp.500–515). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (1994). Strategies of inquiry. In N. K. Denzin & Y.S. Lincoln (Eds.), *Handbook of qualitative research* (pp.199-208). Thousand Oaks, CA: Sage.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21-42). New York, NY: Springer.

Dörnyei, Z. (1990). Conceptualizing motivation in foreign language learning. *Language Learning*, *40*(1), 45-78.

Dörnyei, Z., Csizér, K., & Németh, N. (2006). *Motivation, language attitudes and globalisation: A Hungarian perspective* (Vol. 18). Multilingual Matters Limited.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640-647.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual review of psychology*, *53*(1), 109-132.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education*, *22*(3), 201-204.

Entwistle, N. J., & Entwistle, A. (1991). Contrasting forms of understanding for degree examinations: The student experience and its implications. *Higher education*, *22*(3), 205-227.

Education Testing Service. (2010). Who takes the TOEFL Test? Retrieved March 2, from http://www.ets.org/toefl/ibt/about?WT.ac=toeflhome_ibtabout2_121127

Fish, J. (1988). *Responses to mandated standardized testing*. (Unpublished doctoral dissertation), University of California, Los Angeles.

Flockton, L., Crooks, T., & Baker, L. (2002). *Social Studies: Assessment Results 2001*. Educational Assessment Research Unit, University of Otago.

Frederiksen, N. (1984).The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193-202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* New York, NY: Routledge.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, *26*(1), 123-144.

Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London: Edward Arnold.

Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second language acquisition. *Canadian Journal of Psychology*, *13*(4), 266-272.

Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second language learning*. Rowley: Newbury House Publishers.

George, D., & Mallery, P. (2007). *SPSS for Windows: A simple guide and reference*. (7th ed.). Needham Hieights, MA: Allyn & Bacon.

Grabe, W. (2002). Narrative and expository macro-genres. In A. Johns (Ed.), *Genre in the classroom: Multiple perspectives* (pp.249-267). Mahwah, NJ: Lawrence Erlbaum.

Green, Alison. (1998). *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge: Cambridge University Press.

Green, Anthony. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge: Cambridge University Press.

Grinnell Jr, R. M., & Unrau, Y. A. (2010). *Social work research and evaluation: Foundations of evidence-based practice* (9th ed.). Oxford: Oxford University Press.

Guba, E. G., & Lincoln, Y. S. (1989*). Fourth generation evaluation*. Newbury Park, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1994). Competing Paradigms in Qualitative Research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp.105-117).Thousand Oaks, CA: Sage Publications.

Guiford, J. P. (1946). New standards for test evaluation. *Educational & Psychological Measurement*, 6, 427-438.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement,1*(1), 1-10.

Guion, R. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, 11, 385-398.

Guion, R. (1998). *Assessment, measurement, prediction for personnel decisions*. Mahwah, NJ: Lawrance Erlbaum.

Gu, X. (2003). The need to develop a test on fats reading in CET. *Teaching English in China*, *26*(2), 2-8.

Gu, X. (2004). *Positive or Negative? An Empirical Study of CET Washback on College English Teaching and Learning in China*. (Unpublished doctoral thesis), Shanghai Jiao Tong University, Shanghai, China.

Gu, X. (2007). The empirical study of CET washback on college English teaching and learning in China. *Journal of Chong Qing University (Social Science Edition), 13*(4), 119-125.

Gu, X., & Guan, X. (2003). CET Yuedu Ceshi yu Daxue Yingyu Jiaocai Yidudu Chouyang Yanjiu (A study on the readability of reading materials in CET reading test and CET textbooks). *Journal of Xi'an International studies University*, *11*(3), 39-41.

Hamp-Lyons, L. (1997). Washback, Impact, and Validity: Ethical Concerns. *Language Testing*, *14*(3), 295-303.

Han, B., Dai, M., & Yang, L. (2004). Problems with College English Test as emerged from a survey. *Foreign Language Teaching and Research, 35* (5), 352-358.

Han, M., & Yang, X. (2001). Educational assessment in China: Lessons from history and future prospects. *Assessment in Education: principles, policy & practice*, *8*(1), 5-10.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York, NY: Newbury House.

Hawkey, R. A. (1982). *Investigation of interrelationships between cognitive / affective and social factors and language learning.* (Unpublished doctoral thesis). Institute of Education, University of London, London.

He, L. (2010). The graduate school entrance English examination. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 145-157). New York, NY: Routledge.

Heaton, J. B. (1988). *Writing English Language Tests*. London: Longman.

Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: an intermediate course with examples in S-plus, R, and SAS*. New York, NY: Springer.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Newbury House Cambridge, MA.

Hou, X., &Wang, W. (2008). A study of college student's attitude towards new CET-4 Listening subtest. *Journal of Xi'an International Studies University, 16*(3), 91-94.

House, E.R. (1980). *Evaluating with validity*. Beverly Hills, CA: SAGE Publications.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis, or, Dogmas die hard. *Educational Researcher*, 17, 10–16.

Howe, K. R. (1992). Getting over the quantitative-qualitative debate. *American Journal of Education*, 100, 236–256.

Howie, S. J. (2003). Language and other background factors affecting secondary pupils' performance in mathematics in South Africa. *African Journal for Mathematics, Science and Technology Education*, 7, 1-20.

Huang, D. (2002). *A preliminary study of CET-4 washback*. (Unpublished master's thesis). Southwest Jiaotong University, China.

Huang, J. (2009). *An investigation into immediate and longitudinal CET washback from students' perspective*. (Unpublished master's thesis). Chongqing University, Chongqing, China.

Hughes, A. (2002). *Testing for language Teachers*. (2nd ed.). Cambridge: Cambridge University Press.

Hughes-Wilhelm, K. (1999). Building an adult Knowledge base: an exploratory study using an expert system. *Applied Linguistics*, *20*(4), 425-459.

Hyland, K. (2002). Genre: Language, context, and literacy. *Annual Review of Applied Linguistics*, *22*(3), 113-135.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

IELTS. (2010). IELTS guide for organizations. Retrieved March 2 from http://www.ielts.org/institutions/about_ielts/what_is_ielts.aspx

Jiang, S. (2009). *A content study of CET reading comprehension* (1996-2007). (Unpublished master's thesis). Chongqing University, Chongqing, China.

Jin, Y. (2000). Backwash effect of CET-SET on teaching EFL in China. *Foreign Language World*, 4: 56-61.

Jin, Y. (2004). The development of the CET. *Foreign Languages in China*, 1: 27-29.

Jin, Y. (2005, July). The National College English Test of China. In L. Hamp-Lyons (Chair), *Big Tests*. Symposium at the annual meeting of the International Association of Applied linguistics, Madison, Wisconsin.

Jin, Y. (2006). On the improvement of test validity and test washback: the CET-4 washback study. *Foreign Language World*, 6, 65-73.

Jin, Y. (2008). Powerful tests, powerless test designers? Challenges facing the college English test. CELEA Journal, *31*(5), 3-11.

Jin, Y. (2010). The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.), *English Language Assessment and the Chinese Learner* (pp. 44-59). New York, NY: Routledge.

Jin, Y., & Wu, J. (1997). An application of introspection in research on testing of reading. *Foreign Language World*, 4, 56-59.

Jin, Y., & Wu, J. (1998). Examining the validity of CET reading comprehension by introspection. *Foreign Language World*, 2, 47-52.

Jin, Y., & Yang, H. (2006). The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum*, *19*(1), 21-36.

Johns, A. M. (Ed.). (2002). *Genre in the classroom: Multiple perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14-26.

Jupp, V. (1996). Documents and critical research. In R. Sapsford & V. Jupp (Eds.), *Data collection and analysis* (pp. 298-316). London: SAGE.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kane, M. (2001).Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31-41.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2, 135-170.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2011). Book review: Language assessment in practice: Developing language assessments and justifying their use in the real world. *Language Testing*, *28*(4), 581-587

Kane, M. (2012). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3-17.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Keeves, J. P. (Ed.) (1997) *Educational research methodology and measurement: An international handbook* (2nd ed.). Oxford: Elsevier Science Ltd.

Keeves, J. P. & Sowden, S. (1992). Analyzing qualitative data. In J.P. Keeves (Ed.). *Methodology and measurement in international educational surveys*. The International Association for Evaluation of Educational Achievement (IEA).

Kerlinger, F. N. (1986) *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart & Winston.

Kinnear, P. R., & Gray, C. D. (2008). *SPSS 15 made simple*. Hove, East Sussex; New York, NY: Psychology Press.

Kunnan, A. (1994). Modeling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. *Language Testing*, *11*(3), 225-52.

Kunnan, A. (1998a). An introduction to structural equation modeling for language. *Language Testing*, *15*(3), 295-332.

Kunnan, A. (1998b). Approaches to validation in language assessment. In A. Kunnan. (Ed.), *Validation in language assessment*: *Selected papers from the 17th Language Testing Research Colloquium* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum.

Kunnan, A. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics*, *19*(1), 235-253.

Kunnan, A. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Vol. 9). Cambridge: Cambridge University Press.

Kunnan, A. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27-48). Cambridge: Cambridge University Press.

Kunnan, A. (2010). Test fairness and Toulmin's structure. *Language Testing, 27*(2), 183-189.

Lado, R. (1961). *Language testing*. New York: McGraw-Hill.

Li. J. (2008). *An investigation into the innovated CET washback from teachers' perspective*. (Unpublished master's thesis). Chongqing University, Chongqing, China.

Li. X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development,* 11, 393–404.

Li, Y.G. (2009). *A content validity study on CET listening comprehension (1996-2007)*. (Unpublished master's thesis). Chongqing University, Chongqing, China.

Lincoln, Y.S., & Guba, E. G.(1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Linn, R. L. (2005). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 28-30.

Liu. R. & Dai, M. (2004). On the reform of college English teaching in China. *Teaching English in China*, *27*(4), 3-8.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement*, 3, 635-694.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lynch, B.K. (1996). *Language program evaluation*. Cambridge: Cambridge University Press.

Lynch, B.K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, *18*(4), 351-372.

Madaus, G. (1988). The influence of testing on the curriculum. In L.N. Tanner (Ed.), *Critical Issues in Curriculum: Eighty-seventh Yearbook of the National Society for the Study of Education* (pp.83-121). Chicago: University of Chicago Press.

Markee, N. (1994a). Curricular innovation: Issues and problems. *Applied Language Learning*, 5, 31-30.

Markus, K. A. (1998). Validity, facts, and values sans closure: Reply to Messick, Reckase, Moss, and Zimmerman. *Social Indicators Research, 45*(1), 73-82.

Marton, F. (1981). Phenomenography—describing conceptions of the world around us. *Instructional science*, *10*(2), 177-200.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. Language Testing, *18*(4), 333-349.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *5*(1), 31-51.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension.* Maiden, MA: Blackwell Publishing.

Mehrens, W. A. (1997). *Flagging test scores: Policy, practice, and research*. Washington, DC: American Psychological Association.

Mehrens, W. A. (2005). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Mertens, D. M. (2003). Mixed methods and the politics of human research: The transformative-emancipatory perspective. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp .135-164). Thousand Oaks, CA: Sage.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.13-103). New York: American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3): 241-256.

Minichiello, V., Aroni, R., Timewell, E., & Alexander, L. (1995). *In depth interviewing: Principles, techniques, analysis* (2nd ed.). Melbourne: Longman.

Ministry of Education. (1999). *Higher Education Law of the People's Republic of China*. Beijing, China: China's Law Publishing House.

Ministry of Education. (1999). *National College English Teaching Syllabus* (Rev. ed.). Shanghai, China: Shanghai Foreign Language Education Press.

Ministry of Education. (2004). *College English Curriculum Requirements (For Trial Implementation).* Shanghai, China: Shanghai Foreign Language Education Press.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379-416.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477-496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3-62.

Moore, T., & Morton, J. (2007). Authenticity in the IELTS Academic Module Writing test: A comparative study of Task 2 items and university assignments. In L. Taylor & P. Falvey (Eds.) *IELTS collected papers: Research in speaking and writing assessment* (pp. 197-248). Cambridge: Cambridge University Press.

Moore, T., Morton, J., & Price, S. (2012). Construct validity in the IELTS academic reading test: A comparison of reading requirements in IELTS test items and in university study. In L. Taylor & C. Weir (Eds.) *IELTS collected papers 2: Research in reading and listening assessment* (Vol. 34, pp. 120-211). Cambridge: Cambridge University Press.

Morrison, K. (1993). *Planning and accomplishing school-centered evaluation*. Norfolk: Peter Francis Publishers.

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.). *Innovations in language testing*. London: NEFR/ Nelson.

Moss, P. (1992). Shifting conceptions of validity in educational Measurement: Implications for assessment. *Review of Educational Research*, *62*(3), 229-258.

National College English Testing Committee. (2006). *CET-4 Test Syllabus and Sample Test Paper* (Rev.ed.). Shanghai, China: Shanghai Foreign Language Education Press.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199-215.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3-9.

Noble, A. J., & Smith, M. L. (1994). *Measurement-driven reform: research on policy, practice, repercussion*. CSE Technical Report 381, Tempe, AZ: Arizona State University, CSE.

Oller, J. (1979). *Language tests at school*. London: Longman.

Onwuegbuzie, A. J., & Johnson, R. B. (2004). Mixed method and mixed model research. In R. Johnson & L. Christensen (Eds.), *Educational research: quantitative, qualitative, and mixed approaches* (pp.408-431). Needham Heights, MA: Allyn & Bacon.

Oppenheim, A. N. (1992) *Questionnaire design, interview and attitude measurement.* London: Continuum.

Osgood, D. W., Johnston, L. D., O'Malley, P. M., & Bachman, J. G. (1988). The generality of deviance in late adolescence and early adulthood. *American Sociological Review*, 53, 81-93.

Oxford, R. (1990). *Language learning strategies: what every teacher should know*. New York: Newbury house.

Oxford, R. L. (1996). *Language learning motivation: Pathways to the new century* (Vol. 11). University of Hawaii Press.

Pearson, I. (1988). Tests as levers for change. In D. Chamberlain & R. Baumgartner (Eds.), *ESP in the classroom: Practice and evaluation*. ELT Documents 128. London: Modern English Publications.

Peng. Y. (2009). *An investigation into immediate and longitudinal CET washback from teachers' perspective*. (Unpublished master's thesis). Chongqing University, Chongqing, China.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(10), 26-56.

Popham, J.M. (1987).The merits of measurement-driven instruction. *Phi Delta Kappa,* 68, 679-682.

Prodromou, L. (1995). The backwash effect: from testing to teaching. *ELT Journal, 49*(1), 13-25.

Punch, K. (2009). *Introduction to research methods in education*. London: Sage Publications.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Qi, L. (2004). *The intended washback effect of the National Matriculation English Test in China: Intentions and reality*. Beijing, China: Foreign Language Teaching and Research Press.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stake test. *Language Testing*, *22*(2), 142-173.

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, *14*(1), 51-74.

Qi, L. (2010). Should proofreading go? Examining the selection function and washback of the proofreading sub-test in the national matriculation English test. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 219-233). New York, NY: Routledge.

Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly*, *24*(3), 427-442.

Sasaki, M. (2000). Effects of cultural schemata on students’ test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17*(1), 85-114.

Saville, N., & Hawkey, R. (2004). The IELTS impact study: Investigating washback on teaching materials. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 73-96). Mahwah, NJ: Lawrence Erlbaum.

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (TOEFL iBT Research Report No. TOEFLiBT-08). Princeton, NJ: Educational Testing Service.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education,* (Vol.19, pp.405-450). Washington, DC: American Educational Research Association.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-13.

Shi, Y. (2000). A report on university students’ motivations for English learning [Daxuesheng Yingyu Xuexi Dongji Diaocha Baogao]. *Foreign Language Teaching*, 4, 8-11

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal*, *76*(4), 513-521.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: The National Foreign Language Center at Johns Hopkins University.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331-345.

Shohamy, E. (2001). *The power of tests: a critical perspective on the uses of language tests*. London: Pearson Education Limited.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, *13*(3), 298-317.

Skehan, P. (1989). *Individual differences in second language learning*, London: Edward Arnold.

Spolsky, B. (1975). Language testing: Art or science. *Paper Presentation. Stuttgart: Fourth AILA International Congress.*

Spolsky, B. (1989). *Conditions for second language learning* (Vol. 14). Oxford: Oxford University Press.

Spolsky, B. (1994). The examination of classroom backwash cycle: Some historical cases. In D. Nunan, V. Berry, & R. Berry (Eds.), *Bringing about change in language education* (pp. 55-66). Hong Kong: University of Hong Kong, Department of Curriculum Studies.

Spolsky, B. (1995). *Measured words: The development of objective testing*. Oxford: Oxford University Press.

State Education Commission. (1985). *College English Teaching Syllabus (For college and university students of science and technology).* Shanghai, China: Shanghai Foreign Language Education Press.

State Education Commission. (1986). *College English Teaching Syllabus (For college and university students of arts and science).* Shanghai, China: Shanghai Foreign Language Education Press.

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing: A resource for teachers and administrators*. Alexandria, VA.: Teachers of English to Speakers of Other Languages, Inc.

Someren, M. W., Barnard, Y. F., & Sandberg, J. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample* (TOEFL iBT Research Report No. TOEFLiBT-04). Princeton, NJ: Educational Testing Service.

Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis* (Vol. 748). Boston, MA: Allyn & Bacon.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches.* Applied Social Research Methods Series (Vol.46). Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Taylor, L. (2001). Revising the IELTS Speaking Test: Developments in test format and task design. *Research Notes*, 5, 2-5.

Taylor, L. B., & Falvey, P. (2007). IELTS collected papers: Research in speaking and writing assessment. Cambridge: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument* (2nd ed.). Cambridge: Cambridge University Press.

Urquhart, A. H & Weir, C. J. (1998). *Reading in a second language: Process, product, and practice*. London: Longman.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Vernon, P. E. (1956). *The measurement of abilities*. London: University of London Press.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334-354.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of Language and Education* (Vol.7, pp. 334-343). Dordrecht: Kluwer Academic Publishers.

Wall, D. (1999). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. (Unpublished doctoral dissertation). Lancaster University.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System, 28*(4), 499-509.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching*. Cambridge: Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lanka impact study. *Language Testing, 10*(1), 41-70.

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe: Phase 1, the baseline study* (TOEFL Monograph No. MS-34). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education, 14*(1), 99-116.

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe: Phase 2, coping with changes* (TOEFL iBT Research Report No. TOEFLiBT-05). Princeton, NJ: Educational Testing Service.

Wang, C. (2010). A critical review of reforms on college English teaching and CET from the perspective of EFL learning [Cong Waiyu Xuexi Jiaodu kan Daxue Yingyu Jiaoxue he Kaoshi de Gaige]. *Foreign Language World*, 1, 17-22.

Wang, Changxi (2010). *The CET-4 test paper collection*. Xue Yuan Publishing House, Beijing, China.

Wang, S. & Wang, H. (2011).On the state of college English teaching in China and its future development. *Foreign Languages in China*, *8*(5), 4-11.

Warden, C. A., & Lin, H. J. (2000). Existence of integrative motivation in an Asian EFL setting. *Foreign Language Annals*, *33*(5), 535-545.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, *13*(3), 318–333.

Weir, C. (1993). *Understanding and developing language tests*. London: Prentice Hall.

Weir, C. (2005). *Language testing and validation: An evidenced based approach*. Hampshire, UK: Palgrave Macmillan.

Weir, C. & Roberts, J. (1994). *Evaluation in ELT*. Oxford: Blackwell publishers.

Wen, Q. (2001). *Applied linguistics: research methods and thesis writing*. Beijing, China: Foreign Language Teaching and Research Press.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary educational psychology*, *25*(1), 68-81.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. Snow and D. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp.75-107). Hillsdale, NJ: Lawrence Erlbaum.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Wilson, N. & McLean, S. (1994). *Questionnaires design: A practical introduction*. Newtown Abbey, Co. Antrim: University of Ulster Press.

Wu, Q. (2005, February 25). On the revision of the CET. Second news conference of the Ministry of Education. Retrieved from http//: www.moe.edu.cn/ .

Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21-44.

Yang, H. (2004). An analysis of the English proficiency of the Chinese students as reflected in the national CET test. *Foreign Language World*, 1, 56-60.

Yang, H., & Weir, C.J. (1998). *The CET validation study*. Shanghai, China: Shanghai Foreign Language Education Press.

Yang, P. (2011). *A washbck study on the similarities and differences between the innovated CET-4 and CET-6 from the Perspective of students and teachers*. (Unpublished master's thesis). Chongqing University, Chongqing, China.

Yang, Z. (2010). *A synchronic and diachronic study of CET washback on college English classroom teaching and learning*. (Unpublished master's thesis). Chongqing University, Chongqing, China.

Zhang, Y. (2008). The outline on the reforms of CET-4 and CET-6 [Guanyu Daxue Yingyu Siliuji Kaoshi Gaige de Zongti Silu]. *Foreign Language World*, 5, 2-4.

Zhao, J. & Cheng, L. (2010). Exploring the relationship between Chinese university students' attitudes towards the College English Test and their test performance. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 219-233). New York, NY: Routledge.

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing*, *25*(3), 408-417.

Zou, S. (1997). *The TEM validation study*. Shanghai, China: Shanghai Foreign Language Education Press.