5-2014

An *a Priori* and *a Posteriori* Analysis of Usefulness of TEM 8 and TOEFL

Yuanjun Qi
Grand Valley State University

An *a Priori* and *a Posteriori* Analysis of Usefulness of TEM 8 and TOEFL

Yuanjun Qi

A Thesis Submitted to the Graduate Faculty of

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfillment of the Requirements

For the Degree of

Master of Education

College of Education

May 2014

Dedication

To my dearest mom and dad:

We are always a happy family

and

let's continue our happiness in the future!

I love you!

Acknowledgments

First and foremost, I must give my highly sincere and grateful thanks to my thesis chair, Prof. Wu. Without his most generous advice, this thesis topic would not have existed. It is he who guided me all the way from the proposal of the topic to every hard but fruitful step afterwards. I admire his strict requirements and his sense of responsibility towards his students. No matter how busy he is, he always welcomes his students' questions and will even devote his rest time to answering students' questions. He is definitely a teacher of a lifetime!

Next, I want to thank all my committee members for their willingness to serve on my committee and their active participation and patience during the process. Their suggestions helped improve the structure and the wording of my thesis significantly. Prof. Pearson emailed me her feedback on my thesis proposal in advance of my proposal defense when she could not be present. Prof. Bultsma and Prof. Gu both provided constructive suggestions on my thesis proposal.

Also, I am very thankful for my parents and all my friends who are always by my side when I feel stressful and frustrated. My mother encouraged me almost every day through WeChat. Namrata, my roommate, provided me a peaceful environment in our dorm. Zhao always made me laugh regardless of time and space. Xiaoxin often encouraged me to work hard on my thesis. I spent about five days with Xiaojing, studying together.

Abstract

The present study is dedicated to investigating the construct validity of two large-scale English language tests, Test for English Majors - Band 8 (TEM 8) and Test of English as a Foreign Language (TOEFL). In *a priori* validity research, a qualitative analysis of relevant test specifications and the blueprint of TEM 8 and TOEFL shows that TOEFL has a stronger construct validity evidence than TEM 8. There is overlap in construct definition between TEM 8 and TOEFL. In *a posteriori* validity research, the quantitative analysis of test scores obtained from 48 college junior English majors shows that both TEM 8 and TOEFL share a high item consistency and TEM 8 shares a lower concurrent validity compared with TOEFL.

Content

Chapter One: Introduction

1.1 Problem Statement

This study addresses validity evidence for the construct definition for two large-scale English proficiency tests, namely, Test for English Majors-Band 8 (TEM 8) and Test of English as a Foreign Language (TOEFL). The former is a national test designed by the National Foreign Language Teaching Advisory Board (NFLTA) as an exit exam for Chinese college English majors while the latter is developed by Educational Testing Service (ETS) in the United States and used by North American colleges for admission of international students.

These two large-scale English tests both measure students' English proficiency in areas of listening, speaking, reading, and writing. The biggest difference is that TEM 8 is designed specifically for college senior English majors in China while TOEFL is designed for international students who are applying to North American universities for admission. Despite this difference, there might be overlap in the construct definition, target language use (TLU) domains (Bachman & Palmer, 1996), and assessment task specifications between these two tests.

The China Knowledge Resource Integrated Database (KNS) is recognized as "the most comprehensive gateway of knowledge of China" (China National Knowledge Infrastructure [CNKI], 2010). On KNS, the number of relevant articles to the key words, "TEM 8", "英语专业八级" (the Chinese full name of TEM 8), and "专八" (the Chinese short name), is 254, 226, and 60, respectively. However, after a complete search, only four studies have been found as comparative studies with regard to TEM 8 and other

international large-scale English proficiency tests (Chen, 2010; Xie, 2013; Zhang, 2009; Zhang, 2011).

No research outside Chinese academia has been found in relation to comparison between TEM 8 and TOEFL. Even among those four studies, only one study specifically compares TEM 8, TOEFL, and International English Language Testing System (IELTS). It is a qualitative study without statistical analysis and without much detail, exclusively focusing on comparison of test publishers and test specifications of the writing section. In other words, it lacks indepth and comprehensive analysis.

Existing published research has focused on either of the tests in terms of their usefulness based on the usefulness criteria proposed by Bachman and Palmer (1996), such as construct validity, impact, and reliability, but little has been done to analyze their comparative merits and shortfalls. This lack in our existing knowledge base lends the impetus for this study, as the results of the study will shed light on not only these two large-scale tests compared in terms of their measurement traits, but also provide educational authorities and decision makers with a tool to determine if there is sufficient validity overlap between them to warrant requiring students to take one instead of both.

1.2 Background and Importance of the Problem

TEM 8, initiated in 1991, is the highest-level national certification of English majors in China. A lower level of TEM test is TEM 4 which is designed for sophomore English majors. Developed by the National Foreign Language Teaching Advisory Board (NFLTA), TEM 8 is designed to measure Chinese college students' overall proficiency in English during their senior year. All English majors from all universities/colleges in their senior year must take TEM 8 in order to graduate.

TOEFL, first administered in 1964, is designed by Educational Testing Service (ETS) to measure English proficiency in the skill areas of listening, speaking, reading, and writing of international students who apply for admission to North America schools, primarily in postsecondary educational settings. Different from TEM 8 which is held once a year, TOEFL is held about thirty-five times in mainland China. For example, in 2014, it will be held thirty-eight times ("2014 Schedule," 2013). In 2013, it was held thirty-five times ("2013 Schedule," 2012).

Comparative research on TEM 8 and TOEFL is important for three reasons. Firstly, as high-stakes tests, they are both large-scale English proficiency tests that are of consequence to students: one is an exit exam and the other an entrance exam. Secondly, they both include sections of listening, reading, speaking, and writing, essential constructs in their test specifications. Thus there might be redundancy in the process of testing Chinese English majors' English proficiency. The possibility of redundancy in construct definition and its related validity evidence in testing students' English proficiency should be noted since the beginning of the administration of TEM 8. However, the first published comparative research on these two tests did not emerge until recently (Zhang, 2009). As mentioned earlier, IELTS is also included in this study and it is a qualitative study regarding only the writing section. The comparative research on these two tests needs to be conducted more in depth. They need to be compared comprehensively. For example, the comparison of TLU domains, assessment task specifications, as well as other relevant test qualities need to be conducted.

Thirdly, test takers of TEM 8 are potential test takers of TOEFL. The TOEFL score is accepted by almost all American universities for admission. If Chinese English

majors are interested in pursuing their master's degree in America, they probably need to take the TOEFL. It is obvious that this group of students has to take two English proficiency tests – both TEM 8 to leave a Chinese college and TOEFL to enter an American university. The number of students taking TEM 8 from 1992 to 2010 is shown in Figure 1.1 (Zou, 2010). From Figure 1.1, it is clear that the number of test takers of TEM 8 has been increasing since 1992, and this fact makes the lack of comparative studies on TEM 8 and TOEFL more important.

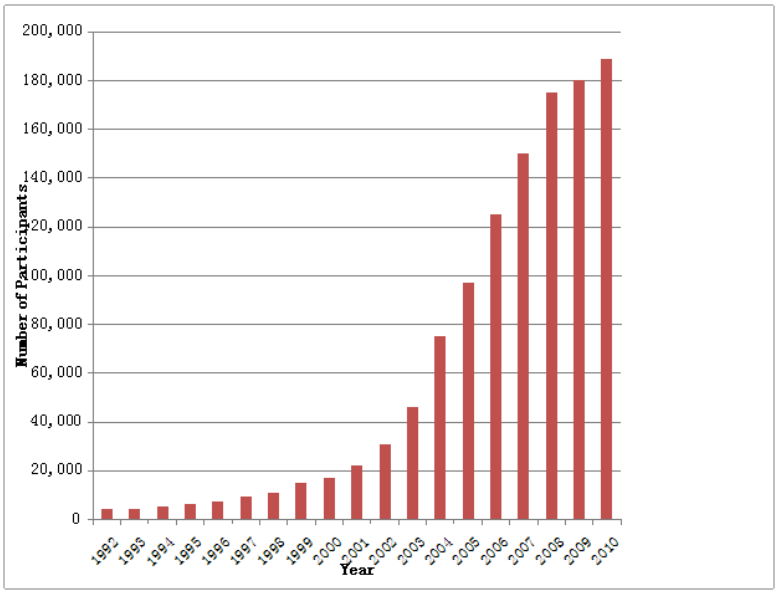
Although the number of participants taking TOEFL is not as detailed as it is for TEM 8, the trend is available through the ETS official website and other websites. For example, according to ETS (2012), "in February, ETS announced a 19 percent increase in the number of Chinese TOEFL iBT test takers for 2011. This represents the largest number of Chinese TOEFL test takers in history" (ETS Reports section, para. 1).

No matter what the conclusion of the comparative research is, it will affect test takers, test developers of both TEM 8 and TOEFL, and test score users. First, it will show test takers the similarities and differences of these two tests, which is beneficial for their future preparation for both tests. Second, it will show test developers of both TEM 8 and TOEFL the strengths and weaknesses of these two tests so that they could better develop subsequent tests. Third, the conclusion of the research will be a useful reference for decision makers in that they could decide whether one could replace the other for entrance admissions, or, to what extent TOEFL can be replaced by TEM 8.

1.3 Research Questions

This thesis attempts to answer the following four research questions. The first two are *a priori* validity questions and the last two are *a posteriori* validity questions:

Figure 1.1 Number of Participants of TEM 8 from 1992 to 2010



1. Do TEM 8 and TOEFL share *a priori* validity evidence in terms of the adequacy of information established for both tests?
2. Do TEM 8 and TOEFL share the underlying construct definition which forms the basis for test specifications and design and which is the most important *a priori* validity?
3. Given the construct definition, different or similar, do TEM 8 and TOEFL share *a posteriori* validity evidence that warrants their use for proclaimed purposes?
4. Because TOEFL is deemed a valid test in measuring international students' English proficiency in terms of the skill areas targeted, to what extent does TEM 8 have concurrent validity, which is *a posteriori* validity, when compared with TOEFL?

1.4 Design, Data Collection and Analysis

The present study is a combination of qualitative research and quantitative research. The former is used for an *a priori* validity study and the latter is used for an *a posteriori* validity study. No personal identifier was collected in the process, so no permission of Grand Valley's Human Research Review Committee (HRRC) was needed.

The data collection site was South-Central University for Nationalities (SCUN), which is located in the central part of China. The participants were 61 voluntary junior English majors. Data sources of the qualitative research came from relevant published documents, which will be described in detail in Chapter Three. Data sources of the quantitative research were test scores obtained from participants who took TEM 8 and TOEFL. Data collection took place outside participants' normal classes at the university.

1.5 Definition of Terms

Assessment task specifications – "provide a detailed description of everything that individual task writers need to know in order to write actual assessment tasks that will support the warrants in the AUA" (Bachman & Palmer, 2010, p. 313).

Authenticity – refers to "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (Bachman & Palmer, 1996, p. 23). Authenticity represents the generalizability of score interpretations.

***A priori* validity evidence** – evidence gathered before test use in support of test design (Weir, 2005, p. 17)

***A posteriori* validity evidence** – evidence gathered after test use in support of post test validation (Weir, 2005, p. 17)

Construct – is "a meaningful interpretation of observed behavior" (Chapelle, 1998, p. 33). For example, in a syntax test, researchers consider test scores as an indicator of test takers' syntax knowledge. Then "syntax knowledge" is the construct in this test.

Construct validity – refers to "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (Bachman & Palmer, 1996, p. 21).

EFL – is short for English as a Foreign Language. For example, English is the foreign language in China. That is, the majority of the Chinese do not use English in their daily life.

ESL – is short for English as a Second Language. For example, English is the second language in India. That is, apart from the first language Hindi, English is also used by Indians in their daily life.

Interactiveness – is "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (Bachman & Palmer, 1996, p. 25). In language testing, individual characteristics mainly include language ability, topical knowledge, and affective schemata.

Large-scale tests – are tests used in the school context. These tests are used to "provide diagnostic information to all stakeholders (teachers, students, parents, school, administrators, etc.), and for state level accountability purposes" (Kunnan, 2008, p. 135). In addition, they are also used in entrance admissions into colleges and universities.

Qualitative research – refers to "the collection, analysis, and interpretation of comprehensive narrative and visual data to gain insights into a particular phenomenon of interest" (Gay, Mills, & Airasian, 2011, p. 630).

Quantitative research – refers to "the collection of numerical data to explain, predict, and/or control phenomena of interest" (Gay, et al., 2011, p. 630).

Stakeholders – refers to "(1) the test developer, (2) the test user, or decision maker, who may also be the test developer, and (3) those individuals, programs, institutions, or organizations that the decision maker and/or test developer specifically targets or intends to be affected by or to benefit from the intended consequences" (Bachman & Palmer, 2010, p. 86).

Target language use (TLU) domain – is "a specific setting outside of the test itself that requires the test taker to perform language use tasks" (Bachman & Palmer, 2010, p. 60).

Usefulness criteria – are criteria used to determine how useful a test is based on the TLL domain the test is designed for: construct validity, reliability, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996, p. 18).

1.6 Delimitations of the Study

Firstly, research questions are about both *a priori* and *a posteriori* validity evidence. Specifically, construct validity and concurrent validity are included. Secondly, research instruments are two real tests. One is the TOEFL adopted from official ETS publication and the other is the 2013 edition of TEM 8. These two tests were chosen because they are the latest tests publicly available. Thirdly, three groups of literature are reviewed in this study. The first two groups are studies of usefulness criteria of TEM 8 and TOEFL. The second group is comparative research between TEM 8 and other large-scale tests.

1.7 Limitations of the Study

Firstly, the study sample in the present research was not randomly selected. Therefore, conclusions from the study have limited generalizability. Random sample might show different results. Secondly, the speaking section of both tests is not addressed in the present research because of constraints, such as the more likely subjective decision of scorers and lack of sufficiently trained scorers. Inclusion of the speaking section would make the results more complete. Thirdly, the literature reviewed for *a priori* validity research of TOEFL is mostly ETS sponsored and most *a priori* validity evidence of TOEFL includes many secondary sources, so the analysis tends to be supportive. Fourthly, the time interval between the use of the two tests was one day, so the practice of the first test might influence the practice of the second test, so the results might be less valid. Fifth, inter-rater reliability was not examined in the present research, so the same scorer might not have scored consistently.

1.8 Organization of the Thesis

In this chapter, general introduction to the present research is presented. Relevant information includes the problem statement, the importance and background of the problem, research questions, the research design, the definition of key terms, and delimitations and limitations of the study. Chapter Two is a literature review which examines independent studies of the usefulness of TEM 8 and TOEFL, as well as comparative studies involving TEM 8 and other large-scale tests. In Chapter Three, the research design is described in detail. For example, participants, instrumentation, data collection, and data analysis will be included. Chapter Four shows the results of the present research. Areas addressed are results of both *a priori* validity research and *a*

posteriori validity research. The last chapter, Chapter Five, is the conclusion of the whole study, which includes a summary of the previous four chapters, conclusions, and discussions, as well as recommendations for future research.

Chapter Two: Literature Review

2.1 Introduction

This chapter provides the conceptual framework and major research regarding usefulness of TEM 8 and TOEFL. Firstly, as conceptual framework for the present study, Assessment Use Argument (AUA) and usefulness criteria, especially concurrent validity, will be briefly introduced.

Secondly, studies of TEM 8 are reviewed in these areas: general background, *a priori* validity evidence, and *a posteriori* validity evidence. The general background consists of three parts: review studies of TEM 8, the change of the syllabus, and the history of validation research of TEM 8. *A priori* validity evidence includes evidence for both the old TEM 8 and the new TEM 8, because *a priori* validity evidence discussed in the literature of the old TEM 8 did not change after the revision of the *National TEM 8 Blueprint* (2004 ed.). *A posteriori* validity evidence is developed in accordance with the six parts of TEM 8: listening comprehension, reading comprehension, general knowledge, proofreading, translation, and writing.

Thirdly, studies of TOEFL are reviewed also in three areas: general background, *a priori* validity evidence, and *a posteriori* validity evidence. In the general background section, history of validation research of TOEFL is briefly introduced. Then, *a priori* validity evidence is discussed regarding five sections: listening, reading, speaking, writing, and other evidence. Finally, *a posteriori* validity evidence contains two parts: speaking and writing prototype, and the whole test. Fourthly, comparative studies between TEM 8 and other large-scale tests are reviewed. Each study is briefly introduced regarding the methodologies and conclusions.

2.2 Conceptual Framework for this Study

The fundamental conceptual framework of the present research is the AUA framework of Bachman and Palmer (2010) and the usefulness criteria proposed by Bachman and Palmer (1996). Specifically, concurrent validity of the usefulness criteria guides one of the research questions, so it will be introduced specifically.

2.2.1 Assessment Use Argument

According to Bachman and Palmer (2010),

> The AUA is a conceptual framework consisting of a series of inferences that link the test taker's performance to a claim about assessment records, to a claim about interpretations, to a claim about decisions, and to a claim about intended consequences, along with warrants and backing to support these claims. (p. 103)

Fundamentally, Bachman and Palmer's (2010) AUA framework takes a comprehensive view of the assessment process from the very initial stages to the final stages. In the initial stage, validity evidence prior to test design is established and in the final stage, completed tests are evaluated empirically for their usefulness based on their six criteria (Bachman & Palmer, 1996). Simply put, every decision made in the assessment process must be analyzed and justified to ensure the test is useful and can stand rebuttals. Rebuttals are "statements that challenge or reject the qualities of the claims" (Bachman & Palmer, 2010, p. 101). This conceptual framework guides this thesis research in that any construct definition pertaining to TEM 8 and TOEFL must have what Bachman and Palmer (2010) term as "backing" and "warrants" and validity evidence must be established for both tests. Warrants are "explicit statements that elaborate one or

more qualities of a claim specifically for the given assessment situation" (Bachman & Palmer, 2010, p. 101). Backing "consists of the evidence that we need to provide to support the warrants in the AUA" (Bachman & Palmer, 2010, p. 102).

Different from the usefulness criteria Bachman and Palmer (1996) proposed, the biggest contribution of the AUA framework is that it links test takers' performance all the way to consequences as a chain. It provides a whole framework which helps test designers effectively design a test. It especially provides guidance for test designers about what they can do before they actually design a test in order to ensure the strongest validity and reliability of a test. The AUA framework guides the first two research questions in that no matter what definition test designers provide, they must have corresponding supporting evidence for that definition. Based on the AUA conceptual framework, the first two research questions were raised to evaluate *a priori* validity evidence of both TEM 8 and TOEFL.

2.2.2 Usefulness Criteria

Bachman and Palmer (1996) proposed six usefulness criteria of a second language test: validity, reliability, authenticity, interactiveness, impact, and practicality. Specifically, the present study focuses on concurrent validity. Based on the construct validity proposed in this usefulness framework, the construct validity of both TEM 8 and TOEFL is examined.

According to Hughes (1989), concurrent validity is used to "see how far results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability" (p.23). Because the present study needs to compare TEM 8 with TOEFL, the concept of concurrent validity is adopted.

Because a large amount of research has been conducted with the sponsorship of ETS to evaluate the validity of TOEFL, in the present study, the TOEFL test is deemed as a valid test and scores test takers earned on TEM 8 will be compared with those on the TOEFL as a way to examine the concurrent validity of TEM 8.

2.3 Studies on the Usefulness of TEM 8

Studies on the usefulness of TEM 8 were reviewed as three groups: general background, *a priori* validity evidence, and *a posteriori* evidence. With the general background, readers would know review studies of TEM 8, the change of the 2004 Blueprint, and the history of validation research of TEM 8. Review of *a priori* validation research and *a posteriori* validation research provides detailed information about what has been done and what needs to be done.

2.3.1 General Background

After being used for twenty-four years, TEM 8 has gone through changes in both the syllabus and the test design. Thus, review studies are divided into four parts: (1) review studies of TEM 8, (2) publication of *the Blueprint* (2004 ed.), (3) the change of the 2004 blueprint, and (4) the history of validation research of TEM 8.

2.3.1.1 Review studies of TEM 8

The latest revision of the test syllabus happened in 2005, so review studies of TEM 8 after 2005 were collected in order to gather the most updated information for the present study. Generally speaking, these review studies emerged mainly after 2010 (Wang, 2013; Zou, 2005; Zou & Chen, 2010; Zou, 2011; Zou, Hong, Zhu, & Zhu, 2012).

According to Wang (2013), in recent years, research on TEM 8 mainly covers six areas. However, due to some similarities among those areas, the researcher of this study

regrouped them into four groups: (1) comprehensive review studies, (2) studies from the perspective of second language assessment, (3) studies from the perspective of other aspects of linguistics, and (4) studies of the interpretation test of TEM 8. With regard to the validity research, Zou (2011) suggested that future validity research on TEM 8 should focus more on test takers because the test will finally be used by test takers and if the test is a valid can also be reflected through the performance of test takers.

2.3.1.2 The change of the 2004 Blueprint

With regard to the change of the 2004 blueprint (Wu, 2005; Zhu, 2005), the most important change is the focus of the test. Although both the old and the new syllabus define TEM 8 as a criterion-referenced test, the focus is changed. In the old syllabus, according to the revision committee of the *Syllabus for TEM 8 (1997)*, TEM 8 is a criterion-referenced test which aims at "testing a single part of and comprehensive English proficiency of test takers" without mentioning the test's relation to English teaching in colleges/universities (p. 1). In the new syllabus, TEM 8 is also defined as a criterion-referenced test, but it aims at checking how the curriculum meets the requirements in the TEM test specifications through testing students who have taken the curriculum. Additionally, according to the revision committee of the syllabus, the new syllabus includes students' grasp of knowledge of the English literature, linguistics, and English societies and cultures. This is a newly-added part in the purpose of the test. The test section corresponding to this purpose is General Knowledge, the requirements of which will be discussed in later sections.

2.3.1.3 The history of validation research of TEM 8

The history of validation research on TEM 8 started from 1993. At that time,

Shanghai International Studies University (SISU) cooperated with the British Council to draw up a three-year plan in order to conduct research on TEM 8's validity and reliability. Data collected from qualitative research and quantitative research were used to verify validity and reliability of TEM 8. Based on the findings, SISU made revisions of the test. The whole project finished in June, 1996.

Validation research of TEM 8 did not stop after 1996. However, it turned from intensive research into phased research. According to Zou and Chen (2010), there are three characteristics of validation research on TEM 8 from 1996 to 2010: (1) more focus on washback effect, (2) continuing research on construct validity, and (3) use of computer scoring technique in order to improve validity. Studies on the washback effect and construct validity mainly come from three sources: published articles, dissertations, and post-doctoral research. The newly adopted theory and instrument include item response theory and structural equation model software. Both qualitative and quantitative research is conducted. Use of computer scoring technique in order to improve validity has been included since March, 2010 at which time computer aided marking has been adopted in scoring the test.

2.3.2 *A Priori* Validity Evidence

A priori validity evidence is of crucial importance before the actual design of TEM 8 in that it determines to what extent the test can reflect the construct definition, checking how the curriculum meets the requirements in the TEM test specifications through testing students who have taken the curriculum. *A priori* validity evidence of TEM 8 is provided corresponding to both the old syllabus (Zou, 1999; Zou, 2003) and the present syllabus (Zou, 2004; Zou, 2005; Zou, 2006; Zou, 2011).

2.3.2.1 A priori validity evidence of the old TEM 8

According to Zou (2003), in order to ensure the validity of TEM 8, the design of the *Syllabus for TEM 8* (1997) is highly dependent on the *National Curriculum for English Majors* (1990 ed.). In addition, test designers paid much attention to the authenticity and interactiveness of the test, since they would affect the validity of the test.

A priori validity evidence of the writing section corresponding to the old syllabus was provided to show how the score of writing section could reflect students' writing abilities to the greatest extent. According to Zou (1999), in designing the writing section, test designers made efforts from mainly six perspectives to ensure the validity of the writing section.

First, a clear purpose of the writing task is given by scenario building. Then, in order to encourage students to think, an assumed viewpoint is given and students are required to express their personal viewpoints towards the topic. Also, assumed target readers are given so that students are aware of the genre they might use. Additionally, rhetorical framework is provided to the students in order to save their time conceiving the organization. Next, exposition and argumentation are tested the most due to their high frequency in students' writing exercise. Finally, topics are selected as familiar as possible to all students in order to reduce the negative influence of topical knowledge. The validity of the above assumed benefit needs to be proved through research.

2.3.2.2 A priori validity evidence of the present TEM 8

According to Zou (2005), the present organization of TEM 8 is an effective reflection of the present *National Curriculum for English Majors* (2000 ed.). For example, it includes the change of time allotment and the number of test tasks in certain sections,

as well as the form and content of test tasks.

Specifically, as an important usefulness criterion which can greatly affect the validity of the test, interactiveness of the listening section was investigated. According to Zou (2004), after making the audio tape of the listening section, test designers, most of whom did not participate in selecting the listening material, would listen to the tape and take notes of the important points they heard. After that, they would compare their notes and select the common ones as test tasks. By doing so, test designers ensure the interactiveness of listening tasks.

In addition, improvement of interactivess of the listening section is also reflected in form and content of test tasks. In order to improve interactiveness, the test requirement of mini-lecture asks students to write no more than three words in each blank. This requirement raises the difficulty of test tasks (Zou, 2004, p. 37). However, the increase of difficulty is a subjective conclusion of Zou (2004). Without empirical evidence, its validity cannot be investigated. At the same time, students need to be highly involved in the listening process in order to determine appropriate answers. The conclusion of the high involvement of test takers also lacks empirical research support. Some students may feel frustrated for the number of words in each blank and they may simply quit listening to the material due to less chance of hearing correct answers.

2.3.3 *A Posteriori* Validity Studies

Most of the validity evidence of TEM 8 is *a posteriori* validity evidence and these *a posteriori* validity studies cover all six sections of TEM 8: listening comprehension, reading comprehension, general knowledge, proofreading, translation, and writing.

In the listening section, test takers listen to a mini-lecture, conversation or an interview, and a news broadcast. Test tasks include gap-filling and multiple choice questions. In the reading section, test takers read four passages containing about 3,000 words and answer altogether 20 multiple choice questions, 5 questions for each reading passage. In the general knowledge section, test takers answer 10 multiple choice questions about English societies and cultures, English literature, and linguistics. In the proofreading section, test takers choose one mistake from each line, and there are ten lines containing about 250 words. The mistakes are about grammar, morphology, rhetoric, etc. In the translation section, test takers translate one passage from Chinese to English and another passage from English to Chinese. Each passage contains 150 words. In the writing section, test takers write about 400 words. Test takers are required to be able to write in any genre, and no possible topics are clarified in the requirement.

2.3.3.1 Listening comprehension

Liu (2010) conducted the content validity research of both TEM 4 and TEM 8 used from 2005 to 2009. She selected five pieces of listening material of TEM 8. The research framework is a revised one based on the original one proposed by Bachman and Palmer (1996). Liu found that the listening material of TEM 8 from 2005 to 2009 shares a high content validity in the following three aspects: input, expected response, and the relationship between input and the expected response. For future research, Liu suggested that listening material from other years should be collected and investigated, as well as other sections of TEM 8. Also, future research can involve test takers in taking the test and analyze test validity from their performance on the test.

2.3.3.2 Reading comprehension

Similar to listening comprehension, the most investigated validity of reading comprehension is content validity (Chen, 2011; Guo, 2012; Jiao, 2008; Lu, 2008; Tian, 2009; Wang, 2009). Based on the general findings of these studies, it is shown that the reading comprehension section of TEM 8 has a high content validity.

Wang (2009) studied reading comprehension of TEM 8 from 1997 to 2008. Wang (2009) compared reading sections with requirements in the *National Curriculum* (2000 ed.) and the *Blueprint* (2004 ed.), and the change of the requirements for the reading section after the revision of the *Blueprint* (2004 ed.). The skimming and scanning part of the reading section was deleted in the *Blueprint* (2004 ed.). Test takers are required to be able to adjust reading speed and reading skills. The number of tasks were reduced from 25 to 20.

Wang (2009) proposed a revised framework based on the framework of task characteristics proposed by Bachman and Palmer (1996). This revised framework includes the setting, the test rubrics, the input, the expected response, and the relationship between the input and the expected response. Based on this framework, the analysis of reading passages and question items from 1997 to 2008 shows that the reading section of TEM 8 has a high content validity.

Guo (2012) examined reading comprehension of TEM 8 in 2010 and 2011. She also proposed a revised framework of task characteristics based on the original one proposed by Bachman and Palmer (1996). This revised framework of task characteristics includes the setting, rubrics, characteristics of input, and the expected response. Guo also compared the reading sections with the requirements in the *National Curriculum* (2000

ed.) and the *Blueprint* (2004 ed.). The result shows that reading sections of TEM 8 in 2010 and 2011 have a high content validity.

Wang (2009) and Guo (2012) both proposed their revised framework of task characteristics, even though they stressed similar components, such as the setting, the rubrics, the input, and the expected response. Under the guidance of the revised framework, they examined the content validity of the reading comprehension section of TEM 8 comprehensively. In contrast, Chen (2011), Jiao (2008), Lu (2008), and Tian (2009) analyzed the reading comprehension section of TEM 8 in different years without a complete framework. The drawback is that their analyses are superficial, but their findings can still be reference for other researchers.

Chen (2011) analyzed the reading comprehension section of TEM 8 in 2010. She analyzed topics, the genre, and the length of reading passages, as well as reading strategies which might be used to answer test items. The result shows that the reading comprehension section of TEM 8 in 2010 has a high content validity.

Jiao (2008) analyzed reading passages from 2005 to 2008. She discussed the genre, topics, and reading abilities tested. Through comparison between reading passages from 2005 to 2008 with requirements in the *Blueprint* (2004 ed.) and the *National Curriculum* (2000 ed.), the researcher concludes that the reading comprehension section from 2005 to 2008 has a high content validity.

Lu (2008) analyzed the reading comprehension section from 2002 to 2006. The researcher examined the genre, topics, and the relationship between the characteristics of reading passages and test items and requirements in the *National Curriculum* (2000 ed.) and the *Blueprint* (2004 ed.). The result shows that the addressed genre and topics are a

small part of the required genre and topics in the *Blueprint* (2004 ed.). The length of reading passages in 2005 and 2006, being 2240 and 2490 respectively, does not meet the requirement which is about 3000 words in the *Blueprint* (2004 ed.).

Tian (2009) analyzed the reading comprehension section of TEM 8 from 2005 to 2007. The length, the genre, topics, and the reading abilities tested are examined. The result shows that the reading comprehension section of TEM 8 from 2005 to 2007 shares a high content validity.

2.3.3.3 General Knowledge

Validation research of the general knowledge section covers mainly construct validity and content validity. According to Zou (2007) and Zou et al. (2009), test tasks in the general knowledge section meet the requirements of the *National Curriculum* (2000 ed.); that is, they cover the three areas required: knowledge of English societies and cultures, knowledge of English literature, and knowledge of linguistics. However, there are no detailed requirements within the above three areas which should be a clear guidance for test designers to design a valid general knowledge section.

On the contrary, Wang (2006) and Wang and Liu (2007) found lack of support of the content validity. Specifically, Wang (2006) used factor analysis to investigate the dimensions of test tasks in the general knowledge section in 2005. He found that TEM 8 in 2005 shared two dimensions in the construct of the general knowledge section: (1) knowledge of linguistics and English literature and (2) knowledge of geography and politics. Additionally, Wang and Liu (2007) also used factor analysis to examine the construct of test tasks in the general knowledge section from 2005 to 2007. They found that test tasks focused more on American history, British novels, pragmatics, and

semantics. In other words, test tasks in the general knowledge section from 2005 to 2007 did not cover all the areas which should be included according to the requirements: knowledge of English societies and cultures, knowledge of English literature, and knowledge of linguistics.

2.3.3.4 Proofreading

Different research has proved that the proofreading section has a high content validity (Han, 2007; Liu, 2010; Lou, 2007; Zou, 2012). Altogether, the examined proofreading section ranges from 1999 to 2012, so the common findings are highly persuasive. Liu (2010) suggested the use of corpus to improve its content validity. Zou (2012) suggested improving the length of the passage and the amount of vocabulary, as well as more genre and topics included.

2.3.3.5 Translation

Only the construct and content validity of the translation section has been investigated. Through inter-correlation analysis, Zhou (2008) proved that this section has satisfactory construct validity. Also, this section meets the basic requirements of the *Blueprint* (2004 ed.), so it also has a satisfactory content validity. However, Zhou (2008) suggested that the instruction should be more comprehensive, such as the author, source, and potential readers included. Additionally, the translation passages from 2005 to 2008 are a little longer than is required in the *Blueprint* (2004 ed.). Also, more topics should be covered. To add the author and source is considered reasonable, but to add potential readers is deemed as unnecessary. It is because that test takers are the potential readers of the passages.

2.3.3.6 Writing

With regard to writing, only the theory-based validity has been investigated. Theory-based validity is part of the construct validity. Xiu (2008) examined the theory-based validity of the writing section of TEM 8 in 2004 through a survey among English senior students from three universities right after their taking the test. Xiu proposed a framework of theory-based validity of writing tests. This framework includes six mental stages in the writing process and two types of knowledge involved in the writing process, namely, the language knowledge and the content knowledge.

The six mental stages include making sure the writing target (what to write), adjusting the topic and genre, brainstorming ideas, organizing ideas, producing expressions, and making revisions. The language knowledge includes grammar knowledge, discourse knowledge, functional knowledge, and sociolinguistic knowledge. The content knowledge includes internal knowledge and external knowledge.

Through factor analysis, it is shown that the six stages test takers went through while writing were basically involved in their taking the writing section of TEM 8. This means that the writing section of TEM 8 reflects students' writing ability. The result of the regression analysis and ANOVA show that the writing section differentiates students' language ability well.

Even though this is the only study that has been found regarding the construct validity of the writing section of TEM 8, it is carefully designed and conducted with its own proposed framework of theory-based validity of writing tests. This study involved test takers in the survey and test takers took the survey right after they finished the writing section. This short interval between the process of writing and the survey

enhances the validity of the results.

2.4 Studies on the Usefulness of TOEFL

This section includes the general background, *a priori* validity evidence, and *a posteriori* validity evidence. In the general background section, the basic history of validation research of TOEFL was reviewed. The *a priori* validity evidence provides literature with regard to each test section in TOEFL. The *a posteriori* validity evidence covers literature about the speaking and writing prototype and the TOEFL test as a whole.

2.4.1 General Background

ETS (2008) presented an overview of research evidence with regard to the validity of TOEFL. They put up six propositions relevant to validity evidence of TOEFL and listed research evidence correspondingly. These propositions addressed content validity, scoring validity, construct validity, and consequential validity.

With regard to the content validity, Taylor and Angelis (as cited in ETS, 2008) and Jamieson, Eignor, Grabe, and Kunnan (as cited in ETS, 2008) reviewed research about English language skills needed at English-medium universities/colleges and summarized frameworks developed for a new test design. Rosenfeld, Leung, and Oltman (as cited in ETS, 2008) surveyed undergraduate and graduate faculty and students and proposed the importance of a variety of English skills for academic success.

In terms of the scoring validity, Brown, Iwashita, and McNamara (as cited in ETS, 2008) and Cumming et al. (as cited in ETS, 2008) examined raters' cognitive processes as they scored test takers' responses. The knowledge of raters' cognitive processes contributes to the design of the scoring rubric.

Construct validity of TOEFL is examined in combination of self-assessment

(Wang, Eignor, & Enright, as cited in ETS, 2008), academic placement (Wang, Eignor, & Enright, as cited in ETS, 2008), local instructional tests for international teaching assistants (Xi, as cited in ETS, 2008), and performance on simulated academic listening tasks (Sawaki & Nissan, as cited in ETS, 2008). The high correlation between TOEFL scores and each of the four components listed above proves the construct validity of TOEFL.

In order to maximize the positive consequence of the test use, ETS (as cited in ETS, 2008) provides guidance on how to set standards for score use for admissions. ETS (2004 as cited in ETS, 2008) also published a manual for English teachers, academic directors, and curriculum coordinators. This manual provides sample tasks and examples of classroom activities in order to promote communicative approaches. Because the intention of the design of TOEFL is that high scores in TOEFL indicate that students can communicate fluently in English-medium universities/colleges, this manual aims at help students truly improve their academic English skills. As a result, students with high English skills will earn high scores in TOEFL.

Xi (as cited in ETS, 2008) examined the effectiveness of speaking scores in decision making about international teaching assistants (ITAs). In this study, the concurrent validity of the speaking section of TOEFL compared with a local ITA-screening test was investigated. The result is that the speaking section of TOEFL has a strong concurrent validity for the use of IAT screening.

Although studies corresponding to the six propositions are presented without much detail, these studies are valuable sources of mainly positive validity evidence. They provide useful information for other researchers on research methods and research

findings.

A possible caveat is the fact that much of this research is sponsored by ETS, so it is predictable that research findings are likely to provide confirmative evidence for the validity of TOEFL scores. Because of this concern, research reviewed is mainly research without the sponsorship of ETS, except for *a priori* validity evidence.

2.4.2 *A Priori* Validity Evidence

Before the actual design of TOEFL tests, preparation was done in order to ensure its *a priori* validity. ETS conducted various *a priori* research addressing the four sections, listening, reading, speaking, and writing. It also examined other related areas, such as a corpus developed for the listening and reading sections and the development of the Committee of the Examiners (COE) model. Thus, *a priori* validity evidence is discussed as five sections: listening, reading, speaking, writing, and other evidence.

2.4.2.1 Listening

Carrell, Dunkel, and Mollaun (2002) studied the effect of four factors, including note taking, lecture length, topic, and two aptitude variables, on 234 ESL students. The two aptitude variables are listening comprehension proficiency and short-term memory. Altogether, 234 ESL students were chosen from five American universities and took the listening comprehension test online. Other instruments included a short-term memory test, the listening test from a paper-and-pencil TOEFL, and questionnaires. The findings confirmed the interactions among note taking, topic, and lecture length. However, in the researcher's opinion, the participants were ESL students at five American universities. They were already in an English-speaking environment, so they cannot be regarded as representatives of TOEFL test takers. TOEFL test takers are international students whose

first language is not English and these test takers are all over the world. Most of them were not in English-speaking countries.

2.4.2.2 Reading

Hudson (1996) discussed issues relevant to the assessment of academic reading from a communicative proficiency perspective. Four competence areas were pointed out as necessary knowledge for academic reading. They were grammatical competence, organizational competence, illocutionary competence, and pragmatic competence. The competence listed above can be used for explaining the differences of test takers' reading performance.

Hudson concluded that, for reading assessment of the TOEFL 2000 project, the following areas can be improved: (1) constructed-response tasks and performance-type tasks need to be included, (2) selected-response formats could be combined with constructed-response tasks, (3) authentic academic tasks, such as reading journal articles and note taking, could be included in reading assessment, (4) thematically related texts need to be promoted and more adaptive texts could be included, (5) a descriptive scoring scale needs to be developed, and (6) reading could be combined with other language skills, such as listening and writing.

According to Hudson, if the above six areas were improved, the construct validity would be enhanced. It is because that the variation of test tasks would improve the interactiveness than mostly multiple choice questions. It is also because that inclusion of more authentic academic tasks would improve the authenticity of the reading section. Another reason is that combination between reading and other language skills would be more similar with real reading tasks students meet in academic settings. All of the above

efforts would contribute to the improvement of construct validity.

2.4.2.3 Speaking

Douglas and Smith (1997) reviewed theories of communicative competence, sociolinguistic and discourse theories, as well as influence of test methods on test performance. Hymes (as cited in Douglas & Smith, 1997) clarified that competence involves both knowledge and ability. Based on this clarification, Bachman and Palmer (1996) proposed two components of communicative language ability - language knowledge and strategic competence.

With regard to sociolinguistic and discourse theories, based on the speech-act theory proposed by Austin (as discussed in Douglas & Smith, 1997) and developed by Searle (as discussed in Douglas & Smith, 1997), Ek (as discussed in Douglas & Smith, 1997) and Wilkins (as discussed in Douglas & Smith, 1997) proposed the notional/functional syllabus which offers a method for teaching second language according to the "functional (speech act) and notional (topic) needs of language learners" (Douglas & Smith, 1997, p. 10).

Brown and Yule (as discussed in Douglas & Smith, 1997) proposed that coherence exists while a person is listening. It is impossible for the speaker to make every piece of information absolutely clear to the listener. Thus the listener needs to fill in gaps between pieces of information. Halliday and Hasan (as discussed in Douglas & Smith, 1997) proposed that cohesion in a spoken situation indicates that interpretation of one component is dependent upon another component.

Brown and Yule recommended that, the design and scoring rubric of the revised speaking test takes the above theories into consideration. For example, based on theories

of communicative competence, the test designers would focus on linguistic competence, textual competence, functional competence, sociolinguistic competence, and strategic competence.

Douglas (1997) reviewed a psycholinguistic model of speech production and methods of testing spoken language. Nine recommendations are made for TOEFL 2000. For example, it is suggested that speaking and listening skills could be tested together and a single aural/oral score is provided.

2.4.2.4 Writing

Hamp-Lyons and Kroll (1997) discussed the nature of academic writing, assessment variations, and test variables, in order to provide insights about writing assessment in TOEFL. With regard to the nature of academic writing, Hale et al. (as discussed in Hamp-Lyons & Kroll, 1997) conducted a survey of academic writing prompts. This study suggested a classification scheme for identifying the nature of writing tasks required of students in 8 disciplines with a heavy enrollment of ESL students. There are 5 categories in this classification scheme. In the discussion of genre, as one of the classifications, Hamp-Lyons and Kroll (1997) considered that it is impossible to include too much genre in the writing section and not all genre is appropriate to be adopted in the test. This is one example of the discussion.

Based on the overall discussion, writing is recommended to be perceived as discourse competence. In other words, writing happens in a certain context, for a certain purpose, and for an intended audience. In order to enhance construct validity, Hamp-Lyons and Kroll (1997) suggested that test designers of the writing section consider modalities (print, oral, visual), rhetorical specifications, the wording of the prompt and

instructions, the subject matter for the test question, and the level of cognitive demand required for test takers. This study was conducted 17 years ago, so these suggestions can be regarded as part of *a priori* validity evidence of the writing section.

2.4.2.5 Other evidence

Biber et al. (2004) described the design and analysis of the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus. Areas addressed include: (1) collection of tests for the T2K-SWAL Corpus, (2) transcription, scanning, and editing of texts in the Corpus, (3) grammatical tagging and tag-editing, (4) analytical procedures, (5) linguistic analyses, and (6) diagnostic tools and resources.

Biber et al. also developed corpus management tools and diagnostic tools. With corpus management tools, users could identify all texts sharing the same characteristics in the Corpus. With diagnostic tools, users know the most important linguistic characteristics of new texts of a particular type. Based on this Corpus, test designers could begin to develop tests that would be a more accurate reflection of language skills test takers would use in university context.

Chapelle, Grabe, and Berns (1997) discussed the formulation of Committee of Examiners (COE) model. The COE model includes the following components: fundamental components for describing language use, how to express the fundamental components effectively, and how to transfer the common understanding of skills to perspectives of applied linguists about communication.

Based on the COE model, the process of developing construct definition is provided for test developers. Firstly, identify and analyze the academic context of interest. Secondly, hypothesize the abilities required in the context. Thirdly, construct relevant

items. Finally, establish a scoring rubric.

Ginther and Grant (1996) reviewed literature regarding the academic needs of native English-speaking college students in the U.S. The literature reviewed involves three groups: (1) literature on examination of the deficiencies of students' abilities or on classification of the writing tasks they must perform, (2) literature that stresses the nature of academic writing tasks concerning the larger theoretical contexts, and (3) literature that stresses what students have done and their perceptions of adapting to the requirements of universities.

This review study shows that the most common assignments are reports, research papers, and critical reviews. Short essays and objective tests are the most common exams. Organization and sentence structure are the greatest weaknesses in students' writing. The obtained information is useful for enhancing the construct validity of TOEFL 2000. The concern for me is that this information is obtained from native English-speaking students, but it will be applied to non-native English-speaking students. This might weaken the construct validity of TOEFL.

2.4.3 *A Posteriori* Validity Evidence

Different from research on TEM 8, independent research on TOEFL is not specifically focused on a single section. Instead, it is about evaluation of the whole test or more than one section. These studies addressed mainly content validity, authenticity, concurrent validity, and predictive validity.

2.4.3.1 Speaking and writing prototype

Cumming, Grant, Mulcahy-Ernt, and Powers (2004) conducted research investigating the content validity and authenticity of speaking and writing prototype tasks

for a new TOEFL. Interviews and questionnaire were used. Participants were seven experienced ESL teachers from three universities. They were asked to rate their students' English proficiency and then review students' performance on the prototype tasks. Then, they provided feedback through interviews and questionnaires. The feedback shows that interpretations of TOEFL scores were mostly (70%) consistent with English levels of students according to teachers' opinions. The findings of this qualitative study are confirmative evidence of content validity and authenticity of the writing and speaking prototypes for a new TOEFL.

2.4.3.2 The whole test

With regard to the complete test, research on TOEFL Internet-Based Test (iBT) addresses areas of predictive validity (Cho & Bridgeman, 2012) and concurrent validity (Kokhan, 2012). Cho and Bridgeman investigated the relationship between TOEFL iBT scores and GPAs of 2594 international students, both undergraduates and graduates. The result of the regression analysis shows that TOEFL shares the predictive validity. This is the first large-scale study investigating the predictive validity of TOEFL involving both undergraduates and graduates.

Kokhan (2012) examined the possibility of adopting TOEFL scores as evidence for decision making. TOEFL is compared with the English Placement Test (EPT), which is an adopted placement test for international students by the University of Illinois at Urbana-Champaign. Through ANOVA, it was found that the correlation between TOEFL and the EPT changes dramatically over time. The correlation between TOEFL and the EPT is stronger when the time interval is short. This correlation decreases dramatically around week 50. After the 50th week, this correlation increases. Thus, it was concluded

that TOEFL lacks evidence of concurrent validity.

2.5 Comparative Studies between TEM 8 and Other Large-Scale Tests

In China, TEM 8 has been compared mainly with IELTS (Xie, 2013; Zhang, 2011), TOEFL (Zhang, 2009), and GRE (Chen, 2010). However, none of the above studies involves the analysis of test scores among tests. What these studies have in common is narrative comparison. Comparison of the tests in these studies is superficial.

2.5.1 Comparison of the listening section

A narrative comparison was made of the listening section between TEM 8 (2005 – 2010) and three published IELTS tests (Xie, 2013). Only the topics and the types of test tasks were compared, so the findings are less persuasive.

2.5.2 Comparison of the reading section

With regard to reading sections, Zhang (2011) analyzed the reading sections of TEM 8 (2006 – 2009) and a published IELTS test. The following areas were compared: (1) characteristics of passage format, (2) the length, (3) topics, and (4) characteristics of test tasks. In the end, Zhang (2011) suggested that TEM 8 could improve from the following aspects: (1) more formats of reading passages involved, (2) more topics involved, (3) more types of test tasks involved, and (4) increasing the number of test tasks and the length of reading passages. Inclusion of more topics would increase the authenticity of the reading section. Increase of the authenticity would contribute to increase of construct validity.

2.5.3 Comparison of the writing section

When comparing the scoring criteria of the writing section of TEM 8 with GRE, Chen (2010) analyzed the different focus of the scoring criteria from four perspectives: (1)

language ability vs. competence of logical analysis, (2) creativity and criticalness vs. normativity, (3) the real-life background vs. test-taking skills, and (4) scoring based on the whole content vs. scoring based on details. Suggestions for the writing section of TEM 8 include: (1) increase the portion of the logical competence in the scoring criteria, and (2) broaden writing topics. Because there is no clear corresponding part of the suggestions in the construct definition of the writing section of TEM 8, it is hard to argue if they will increase the construct validity of the writing section.

A similar study involving TOEFL, IELTS, and TEM 8 also compares the differences of the writing sections. Zhang (2009) analyzed the differences from four perspectives: (1) publisher and construct definition, (2) format of the writing section, (3) characteristics of writing tasks, and (4) scoring criteria. Four suggestions are made for the improvement of the writing section of TEM 8: (1) more writing tasks are needed, (2) scoring report is encouraged, (3) authenticity needs to be improved, and (4) writing topics need to be broadened. The last two suggestions might increase the construct validity of the writing section. Zhang thought that the writing tasks were authentic, they would elicit the same or similar writing skills students use in real situations. Thus the increase of the authenticity will enhance construct validity.

2.6 Summary

This chapter serves as a knowledge base for the present study. Based on the AUA framework, the researcher will examine *a priori* validity evidence for both TEM 8 and TOFEL. Based on the theoretical framework of usefulness criteria, the researcher will investigate the concurrent validity of TEM 8 when compared with TOEFL.

In this chapter, research on the usefulness of each test section of both TEM 8 and

TOEFL was reviewed. Review of studies on the usefulness of TEM 8 and TOEFL, and of comparative studies of TEM 8 and other larger-scale language tests, shows that basically these two tests are both valid tests. Content validity of TEM 8 is the most investigated usefulness criterion. Compared with many independent studies of TEM 8, independent studies of TOEFL are fewer in number, but deeper in depth. Additionally, both groups of test designers conducted *a priori* validity research.

2.7 Conclusion

Review of research on both TEM 8 and TOEFL shows that present literature for both tests covers most types of validity evidence. In addition, comparative research on TEM 8 and other large-scale tests have been conducted mainly by Chinese researchers. That is to say, the comparison on TEM 8 and TOEFL hasn't draw attention from international scholars. Based on the reviewed literature, three types of conclusion can be drawn.

First, review of *a priori* validity research of TEM 8 shows that TEM 8 lacks enough evidence for some propositions. For example, in designing the new listening comprehension section, Zou (2004) stated that by adding the number of words required for each blank, test designers can increase the interactiveness of listening tasks. Also, in designing the writing section of the old TEM 8, Zou (1999) introduced what test designers had done in order to increase the validity of this section. Although most of these points sound reasonable, whether they would really works needs empirical evidence. Without empirical evidence, the above statements by Zou (1999, 2004) may be subjective suggestions.

Second, independent research on TOEFL is much less than independent research

on TEM 8. It is expected that more research of TOEFL without the sponsorship of ETS can emerge that will provide more valuable evidence for the test. As mentioned earlier, the intention of ETS supported research is probably to provide confirmative evidence of TOEFL to persuade all stakeholders that TOEFL is a valid test, and they can trust the interpretations of TOEFL scores.

Finally, comparative research on TEM 8 and TOEFL, as well as other large-scale tests, is superficial. All four published studies are narrative comparison with no statistical analyses of test scores. The four studies mainly address just one section of the involved tests. In addition, the four studies are about *a posteriori* comparison.

Because of reasons detailed above, the present research addresses both *a priori* and *a posteriori* comparison of TEM 8 and TOEFL. Test takers were also involved in the process. Their test scores were used to analyze the extent of *a posteriori* validity evidence for both tests. Except for the speaking section, the rest of the sections was analyzed in order to provide a more comprehensive validity evidence for both tests. The speaking section is eliminated because of constraints, such as lack of competent scorers.

Chapter Three: Methodology

3.1 Introduction

The purpose of the present research is to investigate whether the underlying construct definition of TEM 8 and TOEFL is the same or similar and whether these two tests meet the usefulness criteria conceptualized by Bachman and Palmer (1996, 2010).

Four research questions are adopted to guide the present research: The first question is to examine the adequacy of information established for both tests; the second question is to examine whether TEM 8 and TOEFL share the underlying construct definition; the third question is to determine whether TEM 8 and TOEFL share *a posteriori* validity evidence that warrants their construct definition, and the last question is to determine to what extent TEM 8 has concurrent validity when compared with TOEFL, the proclaimed valid test. In this chapter, four sections are presented: participants, instrumentation, data collection, and data analysis.

3.2 Participants

Participants in the present research included 61 junior English majors at SCUN. Among the volunteers, four were male students and fifty seven were female students. Because eleven test takers only took TEM 8, they were eliminated from the research. Thus, 50 participants took both TEM 8 and TOEFL and were used as a non-randomized sample. In addition, two outliers were picked out by SPSS because of their significant skewing effect that would have rendered a large standard deviation resulting in difficulty in interpreting test scores for validity information. One of the two outliers scored 0 on the writing section of TEM 8 and the other outlier earned the total score of only 7 on TOEFL for unknown reasons. In the end, test scores of forty-eight participants were adopted.

3.3 Instrumentation

Instruments for *a priori* validity research are test task specifications of TEM 8 and TOEFL. Instruments for *a priori* validity research include the *National Curriculum* (2000 ed.), the *Blueprint* (2004 ed.), *National TEM 8 Blueprint – Oral* (2007 ed.), *TOEFL 2000 Framework: A Working Paper*, *TOEFL 2000 Reading Framework: A Working Paper*, *TOEFL 2000 Listening Framework: A Working Paper*, and *TOEFL 2000 Writing Framework: A Working Paper*. The documents for *a priori* validity research were chosen because they provide ample information for the validity research of TEM 8 and TOEFL. Instruments for *a posteriori* validity evidence include TEM 8 in 2013 and a set of TOEFL test. All the documents and tests are publicly available.

3.4 Data Collection

Participants took TEM 8 on Monday evening on March 10th, 2014. They took TOEFL on Tuesday evening on March 11th, 2014. The researcher did not participate in either of the tests. Instead, their teaching secretary proctored both tests.

Participants used paper and pen to finish both tests. Additionally, they did not use headphones to do the listening section due to practicality. The teaching secretary played the mp3 version of the listening section on the computer in front of the classroom and the volume was high enough for each participant. Participants took the reading, listening, and writing sections of the test. Participants wrote their answer directly on test paper instead of a separate answer sheet.

To best imitate the real listening context of the real TOEFL test, the teaching secretary controlled the mp3 material by herself. She played each section and stopped, waiting for participants to write their answer. She asked participants to raise their hands

when they finished. When more than 3/4 participants raised their hands silently, the teaching secretary moved on to the next test item. In the real context, students take this section on computer and they control the answer time by themselves.

Before participants took the tests, the teaching secretary informed them about the confidentiality of their personal information. Also, they were notified that they could quit either test whenever they wanted. In other words, no personal identifier would appear in the present study, and participants were free to quit the test in the middle. In addition, participants were aware of the use of their test scores, that is, for the purpose of a graduate's thesis research.

To ensure scoring validity of both tests, scorers reviewed scoring rubrics for both tests, especially for the translation section and the writing section. All six scorers are graduate students in SCUN. They were chosen by the teaching secretary based on their sense of responsibility and their high GPA.

3.5 Data Analysis

In *a priori* validity research, the documents for both TEM 8 and TOEFL listed in 3.3 were analyzed in order to answer research question one and two. The whole analysis was based on the AUA framework proposed by Bachman and Palmer (2010).

Scores for each section and the overall scores in both tests were collected. Because of the difference between the total raw scores of TEM 8 and TOEFL, being 100 and 120 respectively, the student test scores for both tests were converted as percentages. Scores of listening, reading, and writing sections of both tests were also converted to percentages. For example, if student A's listening score were 15 and the total possible score were 20, his/her listening percentage score would be 15/20*100 = 75. In such a way,

scores for the two tests could be used for comparison.

GVSU's statistical consulting center assisted in the statistical analysis portion of this research: Wilks' Lambda multivariate analysis of variance (MANOVA) to gauge interaction between test scores from the three skill areas (listening, reading, and writing) as a test of overall difference between the two tests, post-hoc t to identify the largest amount of variation among the three variables; Cronbach's alpha to provide information on how consistently test takers answered the test questions as an indicator of item validity, and paired sample correlation coefficients to measure the level of strength of association between non-compensatory composite scores (total scores without considering section score variation) of the two tests.

3.6 Summary

In sum, this chapter introduced the overall research design of the present study, including participants, instrumentation, data collection, and data analysis. Among the sixty-one participants, fifty took both TEM 8 and TOEFL. Instruments include relevant documents and the two tests as listed above. Internal consistency, Wilk's Lambda, paired t-test, and paired sample correlations were used for the data analysis portion of the research.

Chapter Four: Results

4.1 *A Priori* Validity Research

According to Bachman and Palmer (2010), language assessment tasks focus mainly on two types of domains, language teaching domain, and real life domains (p.60). In language teaching domain, language is used for teaching and learning. In real life domains, language is used for purposes other than teaching and learning. Sometimes characteristics of language teaching and learning tasks closely match characteristics of real life tasks. Sometimes they don't match with each other.

Based on the Assessment Use Argument framework proposed by Bachman and Palmer (2010), in order to ensure the meaningfulness and generalizability of interpretations, seven warrants are needed. With regard to relevance to research questions one and two, two warrants were selected to guide *a priori* validity research:

> 1. The definition of the construct is based on a frame of reference such as a course, syllabus, a needs analysis or a current research and/or theory, and clearly distinguishes the construct from other, related constructs.
> 2. The characteristics of the assessment tasks (i.e., setting, input, expected response, types of external interactions) correspond closely to those of target language use tasks. (159-160)

The above two propositions are directly relevant to adequacy of information established for both tests and their underlying construct definition.

4.1.1 *A Priori* Validity Research on TEM 8

According to the Revision Committee of the *Blueprint* (2004 ed.), construct definition is based on the *National Curriculum* (2000 ed.). Thus, before the examination

of construct definition in the *Blueprint* (2004 ed.), it is necessary to examine the corresponding construct definition in the *National Curriculum* (2000 ed.). The construct definition in the *National Curriculum* (2000 ed.), except for pronunciation and use of reference books, are listed in Table 4.1.

Table 4.1 Requirements for Level Eight in the National Curriculum (2000 ed.)

Items	Requirements for Level Eight
Grammar	1. Good command of coherence, such as correlation, ellipsis, and substitution among sentences and paragraphs, 2. Good command of cohesive device in order to coherently express one's ideas.
Listening	1. Be able to understand all types of English conversations in all real situations of communication, 2. Be able to understand special reports covering areas of politics, economy, culture, education, and science and technology from radios and TV stations of English countries, such as CNN, 3. Be able to understand lectures and Q & A sections about the topics listed above, 4. Be able to understand reports of current news on TV and dialogues in miniseries,

Reading	1. Be able to understand editorials and book reviews on British and American newspapers and magazines, 2. Be able to understand historical biographies and literature works of medium difficulty published by English countries, 3. Be able to analyze opinions, ideas, text structure, linguistic features, and figures of speech of above materials, 4. Be able to grasp main ideas and understand facts and details.
Writing	1. Be able to write in any genre, 2. Substantial content, fluent language, appropriate wording, and appropriate expressions,
Translation	1. Be able to apply translation theories and techniques to translate newspaper articles and literature works from English to Chinese, 2. Be able to translate newspaper articles, magazine articles, and general literature works from Chinese to English, 3. Translation needs to be faithful to the

	original texts and fluent in language.
Speaking	1. Be able to communicate fluently and decently with foreign guests about major issues domestic and abroad, 2. Be able to express personal ideas systematically, indepth, and coherently.

As shown above, requirements of Level Eight involve real life domains in the context of English countries. However, English is a foreign language, instead of a second language, in China. Students do not need to use English to communicate outside the classroom. With regards to correspondence between instructional tasks and real-life tasks, Bachman and Palmer (1996) state,

> In cases where there is a close correspondence between the two [instructional tasks and real-life tasks], we can use either or both as a basis for developing test tasks. In cases where there is no obvious real-life domain, or where there is a lack of correspondence between real-life tasks and instructional tasks, the test developer must attempt to design the test tasks in such a way as to balance the qualities of authenticity and impact. (105)

Based on the examination of the requirements listed above and the correspondence of requirements of teaching target and requirements of English classes, some conclusions can be drawn:

Firstly, some guides don't guide. This includes all requirements for listening

teaching, cultural literacy, writing, translation, and speaking. A common shortfall for them is the lack of focus. Although the requirements mentioned above look like containing detailed information about different TLU domains, too many TLU domains amount to no focus as to what is feasible.

With regard to teaching listening, students are required to understand all types of English conversations in all real situations of communication. Regardless of variation of accents, people, not all native English speakers, can use English to communicate about everything. It is almost impossible for students, even teachers, to understand all types of English conversations in all real situations of communication due to variation on register, subject matter, accent, and cultural information. For example, if people use English to talk about laws and students know nothing about law, it is unreasonable to expect them to understand what a "law" conversation is about.

Furthermore, students are required to understand information on politics, economy, culture, education, and science and technology from radios and TV stations of English countries, such as CNN. Additionally, they are expected to understand lectures and Q & A sections about the topics listed above. Also, they are expected to understand reports of current news on TV and dialogues in miniseries. The question is, if a student could understand all the above material, what cannot he/she understand? The second question is, is it possible to help students achieve the above standards by classroom teaching in a country where English is a foreign language? It doesn't really define anything in that it includes almost everything.

Requirements for speaking provide almost no specific information about TLU domains. These requirements lack teachability and measurability, as they include all

possible topical knowledge in both language teaching and real life domains.

Secondly, some of the requirements confuse TLU domains with how these domains are realized. Part of the construct definition includes the test taker's ability to listen to and understand English-medium TV and radio shows. However, TV and radio shows may be used to talk about the same current affairs as newspapers and magazines, but listening to the radio and reading a newspaper about the same news events entail different language skills. By the same token, the single TLU domain of "education" can be diverse, ranging from the language used in classroom teaching to the language used in educational technology, resulting in numerous subdomains which make it impossible to design a language test under such construct definition. Validity evidence gathered from such a broad range of specified and unspecified TLU domains and subdomains is not very useful for test designers.

Requirements for translation also embrace broad TLU domains, so these requirements lack focus. These requirements confuse TLU domains with how these domains are realized, as discussed above. For example, they require students to be able to translate newspaper and magazine articles. Newspaper and magazine articles can possibly include articles covering every real life domain, such as hospitals, courts, shopping centers, etc. In other words, students do not possess topical knowledge about every aspect of life.

Thirdly, if the content of a class is required in the national curriculum, this class should be a class in any school curriculum. Otherwise, not all students will take it. The fact is, though, some students have relevant knowledge and others do not. In the *National Curriculum* (2000 ed.), knowledge from some elective classes is also required in the

teaching requirements. For example, students are required to be able to read editorials and book reviews in British and American newspapers and magazines, historical biographies and literature works of medium difficulty published in English countries. However, the only class about reading newspaper articles, Selected Readings from English Newspapers, is an elective class. Similarly, Selected Readings from English Novels, Selected Readings from English Fiction, Selected Readings from English Plays, and Selected Readings from English Poetry are all elective courses. This course arrangement potentially weakens construct validity claims about reading in the *National Curriculum* (2000 ed.).

In terms of writing, students are required to be able to write in any genre. However, Applied Writing and English News Writing are two elective courses in the *National Curriculum* (2000 ed.). What can be implied from this is that, at least applied writing and English news writing are not required to be taught in relevant compulsory English writing classes. If so, not all students will be able to write in any genre, at least not in applied writing and English news writing.

According to Dai (2010), one of the major problems with current English teaching at Chinese universities/colleges is the discrepancy between their curriculum arrangement and the requirements of the *National Curriculum* (2000 ed.) (p. 5). For example, some universities/colleges weigh curriculum of business or journalism more heavily than the fundamental linguistic and literature classes. Some of them reduce class hours, replace normal classes with lectures, or even cancel those classes. This phenomenon weakens the construct validity of the *National Curriculum* (2000 ed.), and eventually, of the *Blueprint* (2004 ed.). Construct definitions sharing similar

characteristics as discussed above cannot "clearly distinguish the construct from other, related constructs" (Bachman & Palmer, 2010, p.159).

Because most requirements in the *Blueprint* (2004 ed.) and the *Blueprint - Oral* (2007 ed.) keep the original requirements in the *National Curriculum* (2000 ed.), only the different parts will be analyzed here. Table 4.2 shows requirements of the *Blueprint* (2004 ed.) and the *Blueprint - Oral* (2007 ed.).

Table 4.2 Requirements of the Blueprint (2004 ed.) and the Blueprint - Oral (2007 ed.)

Items	Requirements	Selection of Test Material
Listening	1. Be able to understand all types of English conversations in all real situations of communication, 2. Be able to understand special reports covering areas of politics, economy, culture, education, and science and technology from radios and TV stations of English countries, such as CNN, VOA, and BBC, 3. Be able to understand lectures and Q & A sections involving the above topics plus topics of history, linguistics and literature, 4. 150 words per minute, and play	1. Content in mini-lecture is relevant to specialized English courses, 2. Content in conversation is relevant to students' daily life, work, and learning activities, 3. News material from VOA and BBC includes common news reports, brief comments, and speeches, all of which are familiar to students, 4. In principle, vocabulary in the listening material is

	once.	within the range of those required in the *National Curriculum* (2000 ed.).
Reading	1. Be able to understand editorials and book reviews on British and American newspapers and magazines, 2. Be able to understand historical biographies and literature works of medium difficulty published by English countries, 3. Be able to analyze opinions, ideas, text structure, linguistics features, and figures of speech of the above materials; grasp main ideas and understand facts and details; understand literal meaning and underlying meaning; infer and judge, 4. Be able to adjust reading speed and reading skills,	1. Broad topics, including society, science and technology, culture, economy, everyday knowledge, biographies, etc., 2. Broad genre, such as narration, description, exposition, argumentation, advertisement, instructions, charts, etc., 3. Key words are within the range of those required in the *National Curriculum* (2000 ed.).
General knowledge	1. Basic knowledge of geography, history, current situation, cultural	Not Available (NA)

	tradition, etc., 2. Basic knowledge of English literature, 3. Basic knowledge of linguistics.	
Proofreading	1. Be able to recognize and correct mistakes in the passage based on knowledge of grammar, vocabulary, and rhetoric.	NA
Translation	1. Be able to use English-Chinese and Chinese-English translation theories to translate newspaper articles and literature works, 2. Translation needs to be faithful to the original texts and fluent in language.	NA
Writing	1. Be able to write in any genre. 2. Substantial content, fluent language, appropriate wording, and appropriate expressions.	NA
Speaking	1. Be able to interpret from Chinese to English, 2. Be able to interpret from English to Chinese,	1. English – Chinese and Chinese – English interpretation: conversations among English speakers (E –

	3. Make comments on a given topic.	C) and Chinese speakers (C – E). Topics include society, politics, economy, etc., 2. Making a comment: Topics include hot issues in politics, economy, education, science and technology, and society both domestic and abroad

In the listening section, the extra requirement is made with regard to understanding of lectures and Q & A sections and lists of radio stations. Apart from what is required in the *National Curriculum* (2000 ed.), students also need to understand topics of history, linguistics, and literature in lectures and Q & A sections. Additionally, apart from CNN, students are expected to understand special reports addressing reports areas of politics, economy, culture, education, and science and technology on VOA and BBC. The same problem exists for this extra requirement in that it is so broad that it has no focus. Also, if students could understand all the listed contents on CNN, they should be able to understand those on VOA and BBC.

With regard to the selection of test materials in the listening section, there are also some problems. Content in the conversation section is relevant to students' daily life, work, and learning activities. This requirement covers almost every real life domain. Similarly, the requirement for news material from VOA and BBC confuses TLU domains

with how these domains are realized as discussed above. Another shortfall is that it excludes CNN in selection of test material while including it in the test requirement.

In the reading section, the extra requirement is that students should be able to understand literal meaning and underlying meaning and be able to infer and judge. The former requirement is a false statement in that while reading, readers do not separate the literal meaning with the underlying meaning. This distinction exists only for separate words and phrases. With regard to the selection of test material of the reading section, topics might be tested include society, science and technology, culture, economy, everyday knowledge, biographies, etc. This requirement of topics also lacks focus.

According to requirements of the speaking test, not only test takers' speaking proficiency, but also their interpretation skills are tested. There is almost no constraint on speaking topics. The danger is that poor speaking performance might be caused by lack of topical knowledge.

Generally speaking, the biggest problem of the adequacy of information established for TEM 8 is its broad TLU domains and confusion between TLU domains and how these domains are realized. The possible reason and solution of the problem is discussed below.

The fact that English is a foreign language in China may be the reason for the above problems. TEM 8 is a test aimed at checking how the curriculum meets the requirements in the TEM test specifications through testing students who have taken the curriculum, so it belongs to "no specific TLU domain" classified by Bachman and Palmer (2010). "No specific TLU domain" refers to a situation where students are required to learn English because of its importance regarding a country's global economic

development. This suits the geopolitical background of the *National Curriculum* (2000 ed.).

In a situation like this, it is very difficult to identify specific TLU domains for test development. In order to solve problems of this kind, Bachman and Palmer (2010) suggested that,

> In situation as these, we would advise the test developer to attempt to define a TLU domain to which she wants to generalize, based on the attributes of the test takers and the construct to be assessed....If the test developer does not define a reasonable TLU domain, then she will be faced with the problem of having to develop assessment tasks that may have no correspondence to language use outside of the assessment itself, and with not being able to justify generalizing beyond the assessment. (p. 285)

With regard to the situation of English in China, test developers of TEM 8 could conduct large-scale surveys among different stakeholders and gather detailed information about the possible TLU domains and the English ability needed by employers.

4.1.2 *A Priori* Validity Research on TOEFL

According to Jamieson et al. (2000), "the purpose of the TOEFL 2000 test will be to measure the communicative language ability of people whose first language is not English. It will measure examinees' English-language proficiency in situations and tasks reflective of universities in North America" (p. 10).

TOEFL includes three types of subject matter: academic content, class-related content, and extracurricular content. It involves three types of setting: instructional milieu,

academic milieu, and non-academic milieu (Jamieson et al., 2000, p.15).

In the present study, *a priori* validity research is conducted from four aspects: the reading section, the listening section, the speaking section, and the writing section.

The Reading Section. According to Enright et al. (2000), "the reading component of the test will reflect the types of reading that occur in university-level academic settings" (p. 14). The construct definition of the reading section is based on the integration of reading theories. Carver (1997) proposed that a theory of reading involves two types of reading, basic comprehension and reading to learn. Guthrie (1988) considered searching reading as a third type of reading. Perfetti (1997) and Goldman (1997) argued for another type of reading, reading of multiple texts. The purpose of reading of multiple texts is to integrate information across multiple texts. The reading team of TOEFL integrates theories of reading of Carver (as cited in Enright et al., 2000), Guthrie (as cited in Enright et al., 2000), Perfetti (as cited in Enright et al., 2000), and Goldman (as cited in Enright et al., 2000). According to Enright et al. (2000), the integrated construct definition of the reading section includes:

1. Reading to find information (or "search reading"),
2. Reading for basic comprehension,
3. Reading to learn, and
4. Reading to integrate information across multiple texts. (pp. 4-5)

The above types of reading ability can be regarded as four variations of a single reading construct. Based on the construct definition, reading content from any subject area, as long as it is typical of academic studies in university, is considered to be appropriate. Topics covered in the TOEFL test include Arts, Humanities, Social Sciences,

Physical Sciences, and Life Sciences. According to Enright et al. (2000), "it seems appropriate to continue to include as much topic variety as possible in the new test" (p. 16). This is considered appropriate based on the overall construct definition of TOEFL.

A possible drawback of including so many topics is that lack of specialized knowledge may become obstacles for test takers' understanding. In order to ensure impartiality of the reading content, Enright et al. (2000) stated that, "care should be taken to ensure that specialized knowledge of a particular field is not necessary to understand the information presented in the passages" (p. 16). This, before the design of the reading section, ensures impartiality of the test content.

The TOEFL reading team also planned subsequent validity research on the reading construct. They were going to (1) keep collecting literature supporting the construct definition, (2) research how the test is used and the characteristics of test takers, (3) research characteristics of texts for reading assignments, (4) initiate researching the variables controlling the difficulty of current TOEFL tasks and research the variables controlling the difficulty of potential reading-to-learn and reading-to-integrate tasks, and (5) examine the effects of controlling the reading time and/or measuring reading rate. All these efforts will provide persuasive evidence for construct validity of the reading section (Enright et al., 2000, pp. 45-46).

The Listening Section. According to Bejar et al. (2000), "there is a general consensus that no uniformly agreed upon definition exists for listening in either native language studies or second language studies" (p. 2). Even so, the TOEFL listening team still comes up with a general guideline: "developing a listening test that reflects and measures both real-life academic listening and the difficulties that second language

learners encounter" (Bejar et al., 2000, p.4).

With regard to construct definition, the TOEFL listening team decided to adapt the definition from the construct definition in the reading section because they found the lack of research on different effects of various purposes on the difficulty of listening tasks. According to Bejar et al. (2000), the construct definition of the listening section is:

1. Listening for specific information,
2. Listening for basic comprehension,
3. Listening to learn, and
4. Listening to integrate information. (p. 10)

In addition, "listening will be restricted to the types of aural input that students hear in academic situations on North American university campuses" (Bejar et al., 2000, p. 6). Furthermore, possible types of interlocutors in academic listening contexts are classified based on the findings in Power (as cited in Enright et al., 2000), who surveyed faculty members at thirty-four institutions across six disciplines (engineering, psychology, English, chemistry, computer science, and business). Altogether, there are seven groups of interlocutors in academic listening setting. Munby (as cited in Enright et al., 2000) identified eighteen possible social relationships recognized as relevant to an international student and academic listening.

With regard to listening content, three types of content are deemed to be relevant to TOEFL listening section: academic content, class related content, and campus related content. Furthermore, Bejar et al. (2000) clarified that, academic content includes "Life Sciences, Social Sciences, Humanities and Arts, and Physical Sciences"; class related content includes "assignments, due dates, text books, etc"; campus related content

includes “registration, faculty advisor, health care, library help, etc” (p. 45).

In addition, setting/location of the TOEFL listening material is defined as “instructional location, study location, or service location” (Bejar et al., 2000, p. 10). According to Bejar et al. (2000), examples of instructional location are “lecture hall, class, seminar room, laboratory, ect”; examples of study location are “dorm study room, library, instructor’s office, computer center, etc”; examples of service location are “heath center, bookstore, registrar’s office, dining area, business office, faculty advisor’s office, etc” (p. 46).

The Speaking Section. The TOEFL speaking test will measure students’ proficiency of oral communication in academic settings. Even though a large amount of research is about testing general second language oral proficiency, little of it is about testing speaking in academic settings. To date, there is no firm theoretical foundation for constructing a speaking test aiming at testing oral proficiency of international students in academic settings.

Because of this, the TOEFL speaking team develops their own framework about testing oral English in academic settings. According to Butler et al. (2000), the four most typical topics in academic settings include:

1. Academic subjects - the standard content of lectures and texts;
2. Organization of learning activities - discussions of learning strategies and negotiations over procedures and tools;
3. Rules of academic life - largely bureaucratic discourse over the formal requirements of courses and academic regulations; and
4. Daily living events that occur on a campus - service encounters

(bookstore, medical services) and informal discussions with friends. (p. 7)

Based on research findings of Hale et al. (1996), Shaughnessy (1977), D'Angelo (1980) and others, the most used functions in academic settings at the university level are definition, narration, description, comparison and contrast, procedural/process (Butler et al., 2000, p. 9). Additionally, because in the real academic context, students always listen to or read something before they speak or write, the integration of listening, reading, and speaking is considered to have face validity.

The Writing Section. The interest of the TOEFL writing team is "individuals' writing and language abilities rather than their academic knowledge or expressive creativity per se" (Cumming et al., 2000, p. 5). Based on this fundamental belief, three rhetorical functions are recognized by the TOEFL writing team as dimensions of the writing tasks: categorization and analysis, problem-solution, and suasive argumentation. The writing team chose these functions based on "their fundamental integrity to writing in North American university and college contexts, across a range of major academic domains" (Cumming et al., 2000, p. 12).

Based on the rhetorical functions mentioned above, the construct definition of the TOEFL writing section is to "categorize key features and/or analyze and describe relations between them; identify a problem and analyze it and/or propose a solution to it; or state a position, elaborate it, and/or justify it" (Cumming et al., 2000, p. 12). Cumming et al. (2000), based on research findings of Anderson et al. (as cited in Cumming et al., 2000), Chiseri-Strater (as cited in Cumming et al., 2000), Leki & Carson (as cited in Cumming et al., 2000), McCarthy (as cited in Cumming et al., 2000), and Nelson (as cited in Cumming et al., 2000), concluded that "the academic writing that is valued most

by both students and instructors is the writing that contributes directly to course grades" (p. 8).

Based on the above research findings and reviews of needs analysis in English for Academic Purposes (EAP) research (Ginther & Grant, 1996; Waters, 1996), the typical writing genres are recognized as summary writing, experimental (lab) reports, case studies, research papers, and book and article reviews (Cumming et al., 2000, p. 8). Based on research findings of Behrens (as cited in Cumming et al., 2000), Braine (as cited in Cumming et al., 2000), Bridgeman & Carlson (as cited in Cumming et al., 2000), Carson et al. (as cited in Cumming et al., 2000), Eblen (as cited in Cumming et al., 2000), Hale et al. (as cited in Cumming et al., 2000), Kroll (as cited in Cumming et al., 2000), Leki & Carson (as cited in Cumming et al., 2000), Ostler (as cited in Cumming et al., 2000), Sherwood (as cited in Cumming et al., 2000), and Walvoord & McCarthy (as cited in Cumming et al., 2000), it is recognized that "academic writing rarely occurs as an isolated act but rather is often in response to the content of a course" (Cumming et al., 2000, p. 9).

Thus it is clear that the typical writing genres and the writing context recognized by the TOEFL writing team are conclusions based on a large amount of research. The solid research support is persuasive evidence of the construct validity of the writing section. Even if the TOEFL writing team already has the construct definition of the writing section, research of the construct validity needs to continue. Apart from the existing construct definition mentioned above, the TOEFL writing team also planned consequent validity research to gather reactions of potential test users and language experts to prototype tasks in order to refine the construct.

From the above analysis, it is clear that behind the construct definition of TOEFL, there is much research evidence supporting it. This is quite different from the design of the construct definition of TEM 8. Instead of scientific research, designers of the *National Curriculum* (2000 ed.) gathered evidence mainly from various meetings with educational experts and surveys with English graduates and major employers. Especially, surveys with English graduates and major employers focus on considerations which are not directly relevant to the construct definition of the *National Curriculum* (2000 ed.). For example, from publicly available information, these surveys are about whether employers need English talents with only English language knowledge or interdisciplinary talents, employers' satisfaction with English course lineup, postgraduates, and undergraduates, what English graduates lack at work, etc. This is a gap of the evidence of the construct validity in the *National Curriculum* (2000 ed.), and eventually, the *Blueprint* (2004 ed.) and the *Blueprint - Oral* (2007 ed.). It is because that the results of these surveys provide very general feedback on opinions towards the course arrangements and the type of English majors needed by society. These requirements contribute very little to the improvement of the construct validity of TEM 8 which has a very broad construct definition.

In sum, information established for TOEFL is considered as adequate due to its powerful evidence through constant research and improvement as a result of research. The construct definition of TOEFL is very clear and operationalizable. It is clear in that the construct definition for each section is specific as the type of ability measured. A clear statement of types of abilities measured is a useful guidance for test designers.

By contrast, the construct definition of TEM 8 is vague and broad. For example,

requirements of the listening section in TEM 8 address a large number of TLU domains, but they do not provide a clear construct definition for the listening ability that test designers want to measure. In contrast, the construct definition for the listening section in TOEFL is very clear: listening for basic information and basic comprehension; listening to learn and integrate information, which forms a conceptual and operational basis for TOEFL test designers.

4.1.3 The Overlap of the Construct Definition between the Two Tests

There is an amount of overlap of construct definition and TLU domains between TEM 8 and TOEFL, as listed below. The biggest overlap exists in the reading section. The construct definition of the writing section in TEM 8 is much broader than it is in TOEFL. Thus the quality of test takers' writing pieces in TEM 8 is more difficult to judge. No specific construct definition of the listening section and the speaking section is provided for TEM 8. However, from its TLU domains and selection of test materials, we could find overlaps between TEM 8 and TOEFL.

Table 4.3 Similarities of TLU, construct definition and selection of materials between TEM 8 and TOEFL

		TLU	Construct Definition	Selection of Test Material
Listening	TEM 8	1. All types of English conversations in all real situations of communication, 2. Lectures and Q & A sections covering	NA	1. Content in mini-lecture is relevant to specialized English courses, 2. Content in conversation is relevant

		areas of politics, economy, culture, education, science and technology, history, linguistics, and literature		to students' daily life, work, and learning activities,
	TOEFL	1. Real-life academic setting at university level	1. Listening for specific information, 2. Listening for basic comprehension, 3. Listening to learn, 4. Listening to integrate information	1. Academic content, such as Life Sciences, Social Sciences, Humanities and Arts, and Physical Sciences, etc., 2. Class-related content, such as assignments, due dates, text books, etc., 3. Campus related content, such as registration, faculty advisor, health care, library help, etc.
Reading	TEM 8	1. editorials and book reviews in British and American newspapers and magazines, 2. Historical biographies and	1. Be able to analyze opinions, ideas, text structure, linguistic features, and figures of speech, 2. Be able to grasp main	1. Broad topics, including society, science and technology, culture, economy, everyday knowledge, biographies, etc.

		literature works of medium difficulty	ideas and understand facts and details, 3. Be able to understand literal meaning and underlying meaning, 4. Be able to infer and judge.	
	TOEFL	1. Real-life academic setting at university level	1. Reading to find information (or "search reading"), 2. Reading for basic comprehension, 3. Reading to learn,	1. Topics covered include Arts, Humanities, Social Sciences, Physical Sciences, or Life Sciences.
Speaking	TEM 8	1. Every possible real life domain	NA	1. Topics include hot issues in politics, economy, education, science and technology, and society, both domestic and abroad.
	TOEFL	1. Real-life academic setting at university level	1. Be able to make definitions, narrate, describe, compare and contrast, produce, etc.	1. Academic subjects - the standard content of lectures and texts, 2. Organization of learning activities - discussions of learning strategies and

				negotiations over procedures and tools, 3. Rules of academic life - largely bureaucratic discourse over the formal requirements of courses and academic regulations, 4. Daily living events that occur on a campus - service encounters (bookstore, medical services) and informal discussions with friends.
Writing	TEM 8	1. Any genre	1. Substantial content, fluent language, appropriate wording, and appropriate expressions	NA
	TOEFL	1. Real-life academic setting at university level	1. Categorize key features and/or analyze and describe relations between them, 2. Identify a problem and analyze it and/or propose a solution to it,	NA

			3. State a position, elaborate it, and/or justify it.	

4.2 *A Posteriori* Validity Research

A posteriori validity research requires the analysis of performance data by test takers that shows the extent to which a test measures its intended constructs on the basis of the *a priori* validity evidence found to support the development and design of the test.

Descriptive statistics of the student test scores from both tests are shown in Table 4.4:

Table 4.4 Comparisons of Mean Scores, the Standard Deviation, and the Standard Error Mean of TEM 8 and TOEFL

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	listeningP_TEM 8	37.60	48	14.403	2.079
	listeningP_TOEFL	35.83	48	20.986	3.029
Pair 2	readingP_TEM 8	68.96	48	13.951	2.014
	readingP_TOEFL	40.49	48	17.970	2.594
Pair 3	writingP_TEM 8	75.31	48	9.075	1.310
	writingP_TOEFL	73.47	48	14.058	2.029

In the present research, six sections as a whole in TEM 8 are compared with three sections (listening, reading, and writing) in TOEFL. Listening, reading, and writing sections in TEM 8 are compared with the corresponding sections in TOEFL. Research

question three is, "do TEM 8 and TOEFL share *a posteriori* validity evidence that warrants their use for proclaimed purposes?" In order to answer this question, test item consistency of the following test items was calculated: (1) all six sections in TEM 8, (2) the listening, reading, and writing sections in TEM 8, and (3) the listening, reading, and writing sections in TOEFL. The results are shown in Table 4.5:

Table 4.5 Cronbach's Alpha

	Test Item Consistency	Number of Test Sections
TEM 8	.651	6
TEM 8	.515	3
TOEFL	.720	3

The Cronbach's correlation coefficients show a higher level of consistency in test items in TOEFL in measuring its intended constructs than TEM 8. TEM 8 test items show moderate consistency. Thus, the answer to research question three is that both TEM 8 and TOEFL share *a posteriori* validity evidence that warrants their use for proclaimed purposes. At the same time, TOEFL shares higher consistency in test items than TEM 8. *A priori* validity evidence for both tests serves as a basis to ensure that test items consistently measure proclaimed language abilities. To that end, TOEFL does a better job than TEM 8. On the other hand, the differences between test scores of the two tests suggest that the students who performed well on TOEFL did a little worse on TEM 8. The conclusion is that, generally, TEM 8 and TOEFL share *a posteriori* validity evidence that warrants their use for proclaimed purposes, but that shared validity evidence doesn't mean shared construct definition. Each test still maintains its own unique purpose and use:

TEM 8 as an exit exam and TOEFL as an entrance exam, each measuring its own sought-after knowledge and skills in its own construct definition.

The fourth research question is, "does TEM 8 have concurrent validity when compared with TOEFL?" In order to answer this question, Wilk's Lambda multivariate analysis (MANOVA) was used to measure the amount of overall variation among the listening, reading, and writing sections as a whole between TEM 8 and TOEFL. The result is shown in Table 4.6:

Table 4.6 Analysis of Variables across the Listening, Reading, and Writing Sections of TEM 8 and TOEFL

Effect	Value	F	Hypothesis df	Error df	Sig.
Wilks' Lambda	.077	102.724[b]	5.000	43.000	.000

The significant variation with the F-value at $p<0.05$ shows that there is an overall difference between TEM 8 and TOEFL in regards to the skill areas (listening, reading, and writing), which suggests that the two tests tap into different validity areas *a posteriori* given the differences in their construct definition. In addition, paired sample correlations between TEM 8 and TOEFL (Table 4.8) show a low to moderate strength of association between the two tests with regards to the three skill areas, with writing at the bottom.

Therefore, the answer to research question four is that TEM 8 shares comparatively low concurrent validity with TOEFL. They each may validly measure their proclaimed knowledge and skills, and there, in theory, must be shared attributes between

Table 4.8 Paired Sample Correlations of Listening, Reading, and Writing Sections in TEM 8 and TOEFL

		Number of Participants	Correlation	Sig.
Pair 1	listeningP_TEM & listeningP_TOEFL	48	.371	.009
Pair 2	readingP_TEM & readingP_TOEFL	48	.309	.033
Pair 3	writingP_TEM & writingP_TOEFL	48	.141	.338

these knowledge and skills area, yet they still maintain their own brand of identity and can't be substituted for each other in decision making. In other words, it cannot be guaranteed that students who perform well in TOEFL also perform well in TEM 8 and vice versa.

Chapter Five: Conclusion

5.1 Summary of the Study

The present study compares two large-scale English language tests, TEM 8 and TOEFL, which have not been compared systematically on their construct definition to date. Specifically, the following questions are tackled: (1) Do TEM 8 and TOEFL have adequate information, i.e. *a priori* validity evidence, established before the design of both tests? (2) Do TEM 8 and TOEFL share similar or the same construct definition? (3) Do TEM 8 and TOEFL establish *a posteriori* evidence that warrants their proclaimed use? (4) With TOEFL as a valid test, to what extent does TEM 8 share concurrent validity with TOEFL? An *a priori* validity study was conducted to answer the first two research questions and an *a posteriori* validity study was conducted to answer the last two research questions.

In an *a priori* validity study, relevant guidelines for both TEM 8 and TOEFL were examined. They were analyzed under the guidance of the AUA framework proposed by Bachman and Palmer (2010) in that any construct definition pertaining to TEM 8 and TOEFL must have what Bachman and Palmer (2010) term as "backing" and "warrants" and validity evidence must be established for both tests. In an *a posteriori* validity study, test scores obtained from participants taking TEM 8 and TOEFL were used. Test scores for each test were used separately to provide evidence of the adequacy of information established for each test. Test scores of TEM 8 and TOEFL were compared across both tests to gather evidence for the concurrent validity of TEM 8 compared with TOEFL.

5.2 Conclusions

Results show that construct definition established for TEM 8 is less adequate than for TOEFL. Test task specifications of TEM 8 provide very broad TLU domains for test designers. Construct definition of TOEFL has validity evidence from a large number of research findings. Most importantly, there is some, but not a large amount of overlap between TEM 8 and TOEFL in their construct definitions. Some knowledge and skills expected by TOEFL are included in the construct definition of TEM 8. However, some are not. For example, for the reading section, a part of construct definition in TOEFL is that test takers are expected to read to integrate information across multiple texts while TEM 8 does not have this requirement. TEM 8 and TOEFL share different characteristics in terms of *a priori* validity evidence domains, but there is some overlap in construct definition.

Furthermore, results of an *a posteriori* validity study show that TEM 8 and TOEFL share some *a posteriori* validity evidence that warrants their use for proclaimed purposes although those purposes differ. TEM 8 shares lower concurrent validity when compared with TOEFL in that the highest paired section correlation is just 0.371.

5.3 Discussion

5.3.1 Why is there the lack of comparative research between TEM 8 and TOEFL?

The most possible reason is that they are used for different purposes. TEM 8 is used to check how the curriculum meets the requirements in the TEM test specifications through testing students who have taken the curriculum, and TOEFL is used by North American universities as part of evidence to make admission decisions to check if students' English language competence reaches the requirements of college courses. Thus,

it is understandable that scholars might not consider comparing these two tests.

Another possible reason is that they are designed by two different institutions. Each institution only focuses on the usefulness of their own test. For example, all research on the TOEFL official website is about examinations of TOEFL exclusively. Although there is no official website of TEM 8, there is much research investigating various qualities of TEM 8. Most research investigates exclusively TEM 8 itself much like TOEFL research. Even if scholars compare TEM 8 with other large-scale language tests, those studies are superficial, as discussed in Chapter Two.

5.3.2 Why can't the findings provide absolute clear answers to the research questions?

Construct definition and TLU domains provided in the *Blueprint* (2004 ed.) and the *Blueprint* (2007 ed.) for TEM 8 are not clear and specific enough for researchers to analyze exactly what the test measures. This problem might come from the English learning context in China. In China, English is not a second language, but a foreign language. Students have almost no chance to use English outside the classroom, regardless of some activities, such as English corners, which are designated places on a college campus or in a public park where anybody can come and practice speaking English with one another. As far as this researcher knows, topics discussed in English corners are very limited. The most discussed questions are about basic personal information, such as name, major, age, etc.

5.3.3 Why do TOEFL items have more consistency in measurement than TEM 8?

According to findings of the present research, the construct definition for TOEFL is formed on the basis of a large amount of research. This research is more scientific and more persuasive than the foundation for the construct definition for TEM 8 in that

TOEFL developers conduct not only surveys investigating the most required language ability but also research examining how that ability could be most effectively tested. The construct definition of TEM 8 is based on the construct definition of the *National Curriculum* (2000 ed.). The construct definition in the *National Curriculum* (2000 ed.) is not clear, so the construct definition of the blueprint cannot be clear. Some questions regarding the construct validity of the *National Curriculum* (2000 ed.) include: Why does the *National Curriculum* (2000 ed.) require certain language abilities? Why does the *National Curriculum* (2000 ed.) require certain TLU domains?

5.4 Recommendations

Based on the findings of the present research, the following suggestions are proposed for test designers of TEM 8 and the revision committee of the *National Curriculum* (2000 ed.). Firstly, the revision committee of the *National Curriculum* (2000 ed.) could conduct more indepth research to provide evidence for the construct definition in the *National Curriculum* (2000 ed.), the *Blueprint* (2004 ed.), and the *Blueprint - Oral* (2007 ed.). The revision committee of the *National Curriculum* (2000 ed.) needs to make most requirements more specific. Secondly, test designers of TEM 8 could narrow down the scope of the requirements. Present requirements cover a vast array of TLU domains and include too much topical knowledge, especially for the general knowledge section and the translation section. Thirdly, test designers of TEM 8 could adopt the ways TOEFL designers use for describing construct definition, such as “the ability to ...”. By doing so, it is clearer for test developers to design test items. Doing so would likely increase the reliability and validity of TEM 8.

Both TOEFL designers and TEM 8 designers could pay more attention to

comparative research between their test and other large-scale language tests. Both groups of test designers would learn from each other. The expected result is improvement of construct validity, as well as other qualities, of both tests.

Next, English teaching at Chinese universities as a whole needs to clearly define the learning objectives and provide appropriate coursework that reflects the objectives set by the national syllabus. The national syllabus provides a basis for test specifications that lead to test design that measures the corresponding knowledge and skills. Because some universities may not arrange classes as required in the *National Curriculum* (2000 ed.), this fact weakens the validity of TEM 8. Another possible reason for not arranging required classes is the lack of qualified teachers teaching those courses. If that is true, universities need to balance enrollment expansion and recruit more qualified teachers.

Further studies could investigate other usefulness criteria for TEM 8 and compare TEM 8 with other large-scale language tests. Researchers should examine test scores as empirical evidence in support of their studies. For TEM 8, more *a priori* validity research is needed because it is more important to establish a sound basis for test design than collecting test scores after the administration of the test, to find out the lack of validity of the test.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* New York, NY: Oxford University Press.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). TOEFL® 2000 listening framework: A working paper. *TOEFL Report Monograph Series, 19,* 1-60. Retrieved from http://www.ets.org/toefl/research/topics/design/

Biber, D., Conrad, S. M., Randi, R., Pat, D., Marie, H., Victoria, C.,...Alfredo, U. (2004). Representing language use in the university: Analysis of the TOEFL® 2000 spoken and written academic language corpus. *TOEFL Report Monograph Series, 25,* 1-374. Retrieved from http://www.ets.org/toefl/research/topics/design/

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). TOEFL® 2000 speaking framework: A working paper. *TOEFL Report Monograph Series, 20,* 1-28. Retrieved from http://www.ets.org/toefl/research/topics/design/

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* New York, NY: Oxford University Press.

Carrell, P. L., Dunkel, P. A., & Mollaun, P. (2002). The effects of notetaking, lecture length and topic on the listening component of TOEFL® 2000. *TOEFL Report Monograph Series, 23,* 1-64. Retrieved from http://www.ets.org/toefl/research/topics/design/

Chapelle, C. A., Grabe, W., & Berns, M. (1997). Communicative language proficiency: Definition and implications for TOEFL® 2000. *TOEFL Report Monograph Series, 10,* 1-70. Retrieved from http://www.ets.org/toefl/research/topics/design/

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York, NY: Cambridge University Press.

Chen, K. (2010). GRE test and test for English major in China - A comparison of differences on writing section between GRE and TEM-8. *Higher Education Forum, 298,* 94-96.

Chen, H. (2011). Validity analysis on TEM 8 reading comprehension. *Journal of Yancheng Institute of Technology (Social Science Edition), 24*(3), 91-94.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 29*(3), 421-442.

CNKI.NET. (2010). *Introduction.* Retrieved from http://eng.oversea.cnki.net/kns55/support/en/company.aspx

Cumming, A., Kantor, R., Powers, D. E., Santos, T., & Taylor, C. (2000). TOEFL® 2000 writing framework: A working paper. *TOEFL Report Monograph Series, 18,* 1-49. Retrieved from http://www.ets.org/toefl/research/topics/design/

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(2), 107-145.

Dai, W. (2010). A mid-term report of the 4th conference of the Advisory Committee of Foreign Languages Programs in Higher Education. *Foreign Language World, 171,* 2-6.

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. *TOEFL Report Monograph Series, 8,* 1-44. Retrieved from http://www.ets.org/toefl/research/topics/design/

Douglas, D., & Smith, J. (1997). Theoretical underpinnings of the test of spoken English™ revision project. *TOEFL Report Monograph Series, 8,* 1-40. Retrieved from http://www.ets.org/toefl/research/topics/design/

English Committee in the Advisory Committee of Foreign Languages Programs in Higher Education. (2000). *National Curriculum for English majors* (2000 ed.). Beijing: Foreign Language Teaching and Research Press.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). TOEFL® 2000 reading framework: A working paper. *TOEFL Report Monograph Series, 17,* 1-79. Retrieved from http://www.ets.org/toefl/research/topics/design/

ETS. (2008). Validity evidence supporting the interpretation and use of TOEFL iBT™ scores. *TOEFL Research Insight Series, 4,* 1-12. Retrieved from http://www.ets.org/toefl/research/topics/validity/

ETS. (2012, April). ETS reports the largest number of Chinese TOEFL® test takers in history. *ENews Update.* Retrieved from http://www.ets.org/s/toefl/newsletter/2012/19378/ww/index.html

Gay, L. R., Mills, G. E., & Airasian, P. W. (2011). *Educational research: competencies for analysis and applications* (10th ed.). Upper Saddle River, NJ: Pearson Education.

Ginger, A., & Grant, M. (1996). A review of the academic needs of native English-speaking college students in the United States. *TOEFL Report Monograph Series, 1,* 1-42. Retrieved from http://www.ets.org/toefl/research/topics/design/

Guo, W. (2012). *Content validity and certain features of TEM 8 reading comprehension* (Master thesis). Retrieved from China academic journals.

Hamp-Lyons, L., & Kroll, B. (1997). TOEFL® 2000 - writing: Composition, community, and assessment. *TOEFL Report Monograph Series, 5,* 1-44. Retrieved from http://www.ets.org/toefl/research/topics/design/

Han, J. (2007). Validity analysis of proofreading in TEM 8. *Journal of Xinzhou Teachers University, 23*(6), 107-108.

Huang, X., & Wang, J. (2009). A study of the scoring standards in speaking test in TEM 8. *Chinese Translators Journal, 181*, 54-59.

Hudson, T. (1996). Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL® 2000. *TOEFL Report Monograph Series, 4,* 1-24. Retrieved from http://www.ets.org/toefl/research/topics/design/

Hughes, A. (1989). *Testing for language teachers*. New York, NY: Cambridge University Press.

Institute of International Education. (2013, November 12). *Open Doors 2013 - Report on international educational exchange.* Retrieved from http://www.iie.org/en/Research-and-Publications/Open-Doors

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). TOEFL® 2000 framework: A working paper. *TOEFL Report Monograph Series, 16,* 1-73. Retrieved from http://www.ets.org/toefl/research/topics/design/

Jiao, Y. (2008). Analysis on content and construct validity of reading part in TEM 8. *Journal of Shandong Institute of Business and Technology, 22*(6), 121-123.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*(2), 291-308.

Kunnan, A. J. (2008). Large scale language assessments. In N. H. Hornberger (Ed.), *Encyclopedia of Language and Education* (2nd ed.) (Vol. 7, pp. 2275-2295). New York: Springer.

Liu, H. (2010). *A corpus-based study of the content validity assessment of proofreading and error correction in TEM 8* (Master thesis). Retrieved from China academic journals.

Liu, J. (2010). *A content validity study on TEM listening comprehension (2005-2009)* (Master thesis). Retrieved from China academic journals.

Lou, X. (2007). A validity and reliability study of proofreading section in both the old TEM 8 and the new TEM 8. *Journal of Mudanjiang Teachers College (Philosophy Social Sciences Edition), 195,* 46-48.

Lu, X. (2008). Content analysis of reading comprehension of TEM 8. *Journal of China Three Gorges University, 30*(1), 99-101.

New Oriental Online. (2013, July 5). *Schedule for TOEFL 2013 in mainland China.* Retrieved from http://toefl.xdf.cn/201211/9171795.html

New Oriental Online. (2013, December 9). *Schedule for TOEFL 2014 in mainland China.* Retrieved from http://news.koolearn.com/20131219/1001276.html

Revision Committee of the *Syllabus for TEM 8* (1997 ed.). (1997). *National TEM 8* Blueprint (1997 ed.). Shanghai: Shanghai Foreign Language Education Express.

Revision Committee of the *Syllabus for TEM 8* (2004 ed.). (2004). *National TEM 8 Blueprint* (2004 ed.). Shanghai: Shanghai Foreign Language Education Express.

Revision Committee of the *National TEM 8 Blueprint - Oral* (2007 ed.). (2007). *National TEM 8 Blueprint - Oral* (2007 ed.). Shanghai: Shanghai Foreign Language Education Express.

Tian, D. (2008). *A study of content validity of reading comprehension of TEM 8 from 2005 to 2007* (Master thesis). Retrieved from China academic journals.

Wang, S. (2006). An analysis of the psychometric property, construct dimensions, construct invariance and test bias of the general knowledge subtest in TEM 8. *Research in Foreign Language and Literature, 6*(3), 54-63.

Wang, S., & Liu, S. (2007). Construct validation on test items of general knowledge in TEM 8. *Foreign Language Education, 28* (5), 35-39.

Wang, Q. (2009). *A content validity study of TEM reading comprehension (1997-2008)* (Master thesis). Retrieved from China academic journals.

Wang, C. (2013). Review on TEM research in the past decade. *Overseas English, 335,* 113-119.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York, NY: Palgrave Macmillan.

Wu, Y. (2005). Revision and influence of the syllabus for TEM 8. *Foreign Language Teaching Abroad, 97,* 38-42.

Xie, Q. (2013). A comparative study of characteristics of listening comprehension and test-taking skills in TEM 8 and IELTS. *Jiannan Literature (Classics), 358,* 244-246.

Xiu, X. (2008). An investigation into the theory-based validity of essay writing in TEM 8. *Foreign Language Teaching and Research, 312,* 447-453.

Yuloo. (Interviewer) & Lin, L. (Interviewee). (2009). *Changes of TOEFL and IELTS in 30 Years.* Retrieved from Yuloo TOEFL News Website: http://sh.yuloo.com/toefl/news/9110.html

Zhang, K. (2009). A comparative study of writing section in TEM 8, TOEFL, and IELTS. *Kao Shi Zhou Kan, 145,* 5-7.

Zhang, Q. (2011). A comparative study of reading comprehension in TEM 8 and IELTS. *The Guide of Science & Education, 16,* 95-96.

Zhou, Y. (2008). *A validation study of translation in TEM 8* (Master thesis). Retrieved from China academic journals.

Zhu, P. (2005). Changes in the new editions of syllabus for TEM 4 and TEM 8. *Foreign Language World, 151,* 67-79.

Zou, S., Chen, H., & Huang, S. (1996). Data analysis of TEM 4 and TEM 8 in 1995. *Foreign Language World, 61,* 55-61.

Zou, S. (1999). Requirements, problems, and strategies - Evaluation of writing ability in TEM 8. *Foreign Language World, 80,* 57-61.

Zou, S. (2003). The reciprocal relationship between language curriculum and language test - Design and implementation of TEM 8. *Foreign Language World, 144,* 71-78.

Zou, S. (2004). How to achieve interactiveness in listening tests - Designing the new TEM 8 listening subtest. *Media in Foreign Language Instruction, 156,* 33-37.

Zou, S. (2005). Understanding the washback effect of tests - With special reference to the revision of the TEM 4/8 test battery. *Foreign Language World, 155,* 59-66.

Zou, S. (2006). A study of the scientific property of English tests. *Foreign Languages in China, 3*(2), 14-18.

Zou, S. (2007). An investigation into the criterion-referenced nature of the general knowledge component of TEM 8. *Foreign Language World, 123,* 86-94.

Zou, S., Peng, K., & Kong, W. (2009). Exploring the construct validity of the general knowledge section in TEM 8. *Foreign Languages in China, 31,* 45-52.

Zou, S., & Chen, W. (2010). TEM tests: Past, present, and future. *Foreign Language World, 186,* 9-25.

Zou, S. (2011). On enhancing test fairness. *Foreign Language Testing and Teaching, 1,* 42-50.

Zou, S., Fang, X., & Chen, W. (2012). Test report for TEM 4/8-2011K. Foreign *Language Testing and Teaching, 5,* 1-10.

Zou, S., Hong, G., Zhu, Y. S., & Zhu, G. (2012). Inherit the past, usher in the future - Opinions from specialists on TEM 4 and TEM 8. *Foreign Language Testing and Teaching, 8,* 1-13.

Zou, Z. (2012). Validity analysis of proofreading section in TEM 8 from 2005 to 2011. *Journal of Hubei Radio & Television University, 381,* 143-144.