The Impact of TEM-8 (Test for English Majors Band 8) on English Majors in China

Cai Wen

Kristianstad University

School of Teacher Education

English, Spring 2010

Level IV English

Tutor: Lena Ahlin

Table of Contents

1. Introduction

A test is a number of questions or exercises to find out how good someone is at something or how much they know. According to different education aims, test types, test standards and test scorings, testing can be divided into different kinds. In the education area, testing can be used to evaluate education, diagnose learning, and help learning (Chang *et al*. 2006: 18). It is very important in the process of education.

TEM, Test for English Majors, is a particular EFL (English as a foreign language) test in China. It was set up by the State Education Commission in 1991, and has been organized by the Higher Education Institution Foreign Language Major Teaching Supervisory Committee since then. The test has been running for about 20 years. It was set to test the actual performance of *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000). There are two levels in TEM, TEM 4 and TEM-8. TEM-8, Test for English Majors Band 8, is based on a higher level of standard. The object of this test is all English majors when they are in their fourth year as well as their last year in college, or more specifically, in their eighth term, which is why the test is called TEM-8. It mainly tests students' ability to use English as a foreign language in addition to testing students' knowledge of words and grammar.

In terms of measuring students' integrative English language ability, TEM-8 is the hardest test for English majors in China (Li *et al*. 2007: 78). Every year, hundreds of thousands of English majors all over the country attend the test. It is an event that each English major student experience before graduation.

However, testing must have some kind of effect upon the process of education, especially TEM-8, which is a very important test for English majors. Here comes the term of "impact". Wall (1997: 291) defines impact as "any of the effects that tests may have on individuals, policies or practices, within the classroom, the school, the educational system, or society as a whole". From individuals to the society, the impact has a wide educational context. Previous researches have worked on the validity of TEM-8, the authenticity of TEM-8, but seldom the impact of TEM-8.

This study is to investigate the overall impact of TEM-8 on students. This is what the education department and the English majors themselves want to know more about.

1.1 Aim

This study aims to find out TEM-8's impact on English major students, in terms of their daily life, learning process, and future life. Before that, this study first analyzes the system of TEM-8 and its impact on students, including the organization of the test, the implementation of the test, and the reform of the test. In addition, the test itself is also analyzed to measure whether the test is suitable for test takers, mainly focusing on the usefulness of the test. There are six qualities within the usefulness, namely reliability, validity, authenticity, interactiveness, impact, and practicality. However, as the main point of this essay, the impact is especially emphasized. In addition to the students, the impact on teachers, education system and the society is mentioned as well.

1.2 Material

The test material used in this study consists of two parts, the test syllabus as well as a sample test paper. *The Syllabus of Test for English Majors Band 8* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2005) helps to analyze the system of the test as well as its impact on students. In addition, the sample test paper of 2009 is used in this research as well. The TEM-8 test is extremely well protected by the committee, so it is impossible to know the content of the test of the year before or even shortly after the test. After finishing the scoring of the test papers, scorers put the pictures of the test on the internet. This year's test was taken less than two months before this study takes place, so the test has not been put on the internet yet. However, since the tests are similar from year to year, thus the test of 2009 is used in this research to help analyze the usefulness of the test.

Two groups of people are involved as participants in this study. The first group is 50 English major senior students of a university in China. They took the questionnaire about what they think about the test, how they prepared for the test, and how the test influences their life. Since the test is taken in March, these students have just experienced this year's test. Therefore, their feelings and opinions are very important for analyzing the impact of the test on students. The other group

consists of two English major teachers of the same university in China. The two teachers have been teaching courses directly related to the TEM-8 for years. The interviews of them can directly reflect the impact of the test on the teaching structure, as well as teachers' attitude towards the courses and the test.

1.3 Method

Analyzing test material is the first main research method used in this study. *The Syllabus of Test for English Majors Band 8* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2005) includes the aim, the nature, the organizer, the participants, the time, the framework, and the requirement of the test, all of which help to analyze the system of the test as well as its impact on students. In addition, the 2009 TEM-8 test paper is used in this study as well, mainly to analyze the usefulness of the test, especially the reliability of the test.

In addition to the test material, questionnaire is used in the research as well (see Appendix 1). The questionnaire is used to find out what students think about the test, how they prepared for the test, and how the test influences their life. Questions 1 to 8 are about students' opinion about the reliability, validity, authenticity, interactiveness of the test. Questions 9 to 11 investigate how students prepared for the test. The rest questions, namely Question 12 to Question 16 are concerned with how the test influences students' life. The questionnaire was sent to the participants in China via email. After collecting the feedback from the students, the data and information are used to investigate the impact of the test on students.

The third research method used in this study is the interviews with teachers (see Appendix 2). The interviews were also carried out via email. The two teachers who are in charge of the course answered the questions about the settings of the course, including the reason why set up the course, the content of the course, and the timetable of the course. Furthermore, the teachers gave their opinions about the teaching structure and students' preparation for TEM-8. The information collected from the interviews is used to analyze test's impact on teachers as well as the educational system.

1.4 A Brief Introduction of TEM-8

According to State Education Commission's *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000: 4-5 my translation), the teaching task and aim of the basic level (Grade One and Grade Two) of English major in Higher Education Institution is to

> teach English basic knowledge, train students' basic skills comprehensively and strictly, develop students' ability of using English in reality, help students form good learning styles and appropriate learning methods, develop students' abilities of logical thinking and independent work, enrich students' knowledge of society and culture, enhance students' sensitivity to the differences among different cultures, and make the students set the stage for senior grades' study. (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000: 4-5 my translation)

In other words, the syllabus means to make the students of basic level to be competent in every respect of English learning. On the other hand, for senior grades students (Grade Three and Grade Four), the syllabus brings forward higher standard. For those students, they should go on learning basic ability of language. Meanwhile, the students should make further efforts on enlarging their scope of knowledge. The emphasis of this stage is on developing students' integrative competence of English, enriching cultural knowledge, and enhancing the ability of social intercourse.

TEM-8 is just the test to assess and evaluate the actual performance of *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000) for senior students. In the meantime, TEM-8 also can assess the teaching quality as well as the students' language ability, particularly the integrative language ability and communicative ability mentioned in the syllabus that Semester 8 students should have achieved. In this way, the test is able to promote the implementation of the syllabus so as to improve teaching quality.

The nature of TEM-8 is a test that assesses test taker's single and integrative language ability. The language abilities tested in TEM-8 include listening ability, reading ability, writing ability, and translating ability. Since the condition for testing oral ability in large scale is still immature, the supervisory committee has to put off this aspect of testing at this stage.

The test is organized in March every year, usually on the Saturday of the first week. For example, this year, 2010's TEM-8 is carried out on March 7th. During that time, Semester 8 just begins and students return from Spring Festival. The supervisory committee means to assess the English major students' language ability by the end of their college or university life, therefore, they choose this time to run the examination.

The test contains six parts, i.e. Listening Comprehension, Reading Comprehension, General Knowledge, Proofreading and Error Correction, Translation, and Writing. The total time for the test is 195 minutes.

There have been two reforms since the test was set up in 1991, respectively in 1997 and 2004. The new tests both came into effect in the following year, which is in 1998 and 2005. In this essay, the reform of 2004 is discussed here since this reform is the latest one and we are all dealing with this version now. In order not to be confused, the tests between 1998 and 2004 are called the old tests, while the tests from 2005 are called the new tests.

2. Theoretical Background

When analyzing a test, several aspects should be included, such as the purpose of testing and the usefulness of the test. The test usefulness is composed of six test qualities—reliability, validity, authenticity, Interactiveness, impact, and practicality, all of which are discussed in turn in detail below.

2.1 The Purpose of Testing

Testing is divided into different types according to different purpose. There are mainly four types of test -- proficiency tests, achievement tests, diagnostic tests, and placement tests (Hughes, 2003: 11). However, a test can also be a combination of two or more types of test such as the test being analyzed in this essay, which is a combination of proficiency and achievement test.

As Hughes (2003: 11) states, proficiency tests are designed to measure people's ability in a language, regardless of any training they may have had in that language. Since it is to test whether one is proficient or not, the content of the test is not probably based on the content or

objectives of language courses that people taking the test would have. Proficiency tests test people on their command of the language for a particular purpose. They may also show whether candidates have reached a certain standard with respect to a set of specified abilities. However, the preparation of proficiency tests is not that easy. Test designers need to consider as objectively and seriously as possible about the instruction, items, structures, and other aspects of the test. Besides, test examiners need to be objective in scoring as well. The examiners are usually independent of teaching institutions, or randomly chosen from all the teaching institutions to make sure they can make fair comparisons between candidates from different institutions.

Achievement tests are the ones that teachers are more likely to be involved in. In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives (Hughes 2003: 12).

2.2 Test Usefulness

The most important quality of a test is its usefulness, since the most important consideration in designing and developing a language test is the use for which the test is intended (Bachman & Palmer 1996: 17). The test usefulness includes six test qualities—reliability, construct validity, authenticity, interactiveness, impact, and practicality. These six test qualities all contribute to test usefulness, so that they should not be evaluated independently of each other.

2.2.1 Reliability

Reliability is often defined as consistency of measurement in a test, which means reliability can be considered a function of the consistency of scores from one set of tests and test tasks to another (Bachman & Palmer 1996: 19). This can be presented as in Figure 1 when reliability is considered to be a function of consistencies across different sets of test task characteristics.

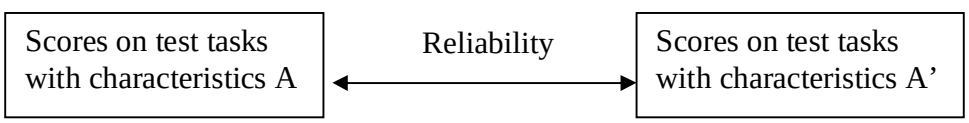

Figure 1: Reliability (Bachman & Palmer 1996: 20)

In this figure, the double-headed arrow is used to indicate a correspondence between two sets of task characteristics (A and A') which differ only in incidental ways.

Consistency, in educational assessment, appears in three varieties—stability, alternate form, and internal consistency. Stability consistency refers to consistency of results among different testing occasions, in other words, consistency over time; alternate-form consistency is about consistency of results among two or more different forms of a test, which is same as equivalence; internal consistency is related to consistency in the way an assessment instrument's items function (Popham 2002: 28).

However, due to the differences in the exact content being assessed on the alternate forms, environmental variables such as fatigue or lighting, or student error in responding, no two tests will consistently produce identical results (Wells & Wollack 2003: 2). This is true regardless of how similar the two tests are. Even the same test administered to the same groups of students but on different occasions will result in different scores. Nevertheless, the students' scores are expected to be similar. The more similar the scores are, the more reliable the test is said to be.

Score Reliability

According to Wells and Wollack (2003: 2), reliability provides a measure of the extent to which an examinee's score reflects random measurement error. One of three factors causes measurement errors:

> (a) examinee-specific factors such as motivation, concentration, fatigue, boredom, momentary lapses of memory, carelessness in marking answers, and luck in guessing, (b) test-specific factors such as the specific set of questions selected for a test, ambiguous or tricky items, and poor directions, and (c) scoring-specific factors such as nonuniform scoring guidelines, carelessness, and counting or computational errors. (Wells & Wollack 2003: 2)

These errors are random and their effect on a student's test score is unpredictable. Sometimes they help students to write the right answer while other times they make students answer incorrectly. Therefore, it is desirable to use tests with good measures of reliability.

Score reliability means if a particular candidate performs in exactly the same way on the two occasions, he would be given the same score on both occasions. In other words, any one scorer

would give the same score on the two occasions, and this would be the same score as would be given by any other scorer on either occasion (Hughes 2003: 43). When scoring requires no judgement, such as the multiple choices test, and could in principle or in practice be carried out by a computer, the test is said to be objective and consistent. Meanwhile, the scorer reliability coefficient is 1, which means the test would be given precisely the same scores for a particular set of candidates regardless by whom or when it happened to be administered. But when a degree of judgement is called for on the part of the scorer, as in the scoring of writing, perfect consistency is not to be expected (Hughes 2003: 43). If so, the scorer reliability coefficient falls below 1.

2.2.2 Validity

According to Messick (1993), validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. In short, validity is the extent to which the instrument measures what it means to measure.

Three Types of Validity Evidence

There are three types of validity evidence—content related, criterion related, and construct related. Content-related evidence of validity refers to the extent to which an assessment procedure adequately represents the content of the assessment domain being sampled; criterion-related evidence of validity is about the degree to which performance on an assessment procedure accurately predicts a student's performance on an external criterion; construct-related evidence of validity refers to the extent to which empirical evidence confirms that an inferred construct exits and that a given assessment procedure is measuring the inferred construct accurately (Popham 2002: 52).

As Hughes (2003: 26) argues, a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned. A specification of the skills or structures, etc. that it is meant to cover is needed for judging whether a test has content validity or not. The greater a test's content validity is, the more likely it is to be an accurate measure of what it is supposed to measure. In addition, a test with low content validity can have a maleficial backwash effect since areas that are not tested are likely to become areas ignored in teaching and learning. Face validity is a component of content

validity. It is established when an individual reviewing the instrument gives the conclusion that it measures the characteristic or trait of interest (Miller 1985: 3). In other words, it looks as if it indeed measures what it is designed to measure.

Criterion-related validity relates to the degree to which results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability (Hughes 2003: 27). This kind of evidence helps educators decide how much confidence can be placed in a score-based inference about a student's status with respect to an assessment domain.

Construct validity has been increasingly used to refer to the general, overarching notion of validity in recent years. It pertains to the meaningfulness and appropriateness of the interpretations that we make based on test scores (Bachman & Palmer 1996: 21). Based on the meaning of a construct in an educational assessment, construct validity is used to refer to the extent to which we can interpret a given score as an indicator of the abilities or constructs we want to measure. We can interpret construct validity as in Figure 2.

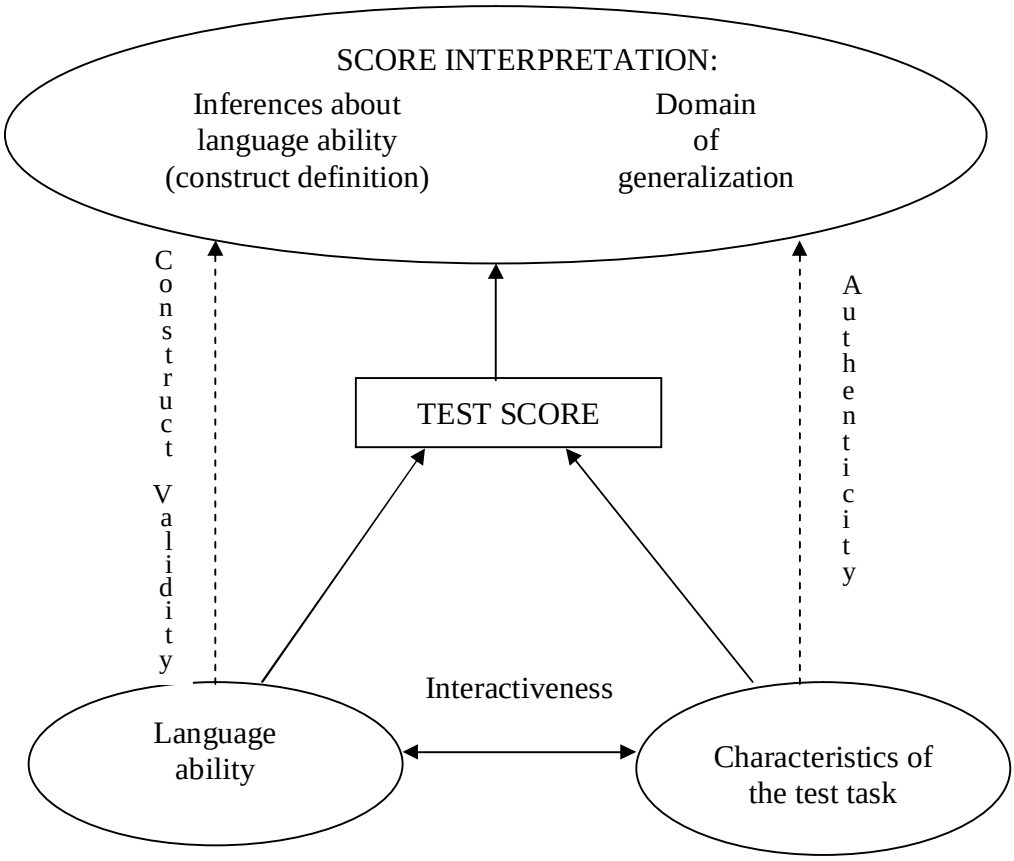

Figure 2: Construct validity of score interpretations (Bachman & Palmer 1996: 22)

Construct validity also has something to do with the specific domain of generalization, construct definition, characteristics of the test task and test taker's areas of language ability.

If a test is to have validity, not only the items but also the way in which the responses are scored must be valid. As Bachman and Palmer (1996: 33) states, if a test is meant to measure more than one ability, it makes the measurement of the one ability in question less accurate. Evidence like content relevance and coverage, concurrent criterion relatedness, and predictive utility can be provided for a particular score interpretation, as part of the validation process. However, valid is just the degree of measurement, and test validation is an on-going process and the interpretations we make of test scores can never be considered absolutely valid (Bachman & Palmer 1996: 22).

How to Make Tests More Valid

In the development of a high stakes test, such as University Entrance Examination, which may have significant effect on candidates' lives, there is an obligation to carry out a valid exercise before the test is taken in operation. Since full validation is unlikely to be possible, as stated above, there are several recommendations from Hughes (2003: 33-34):

> First, write explicit specifications for the test which take account of all that is known about the constructs that are to be measured. Make sure that you include a representative sample of the content of these in the test.
>
> Second, whenever feasible, use direct testing. If for some reason it is decided that indirect testing is necessary, reference should be made to the research literature to confirm that measurement of the relevant underlying constructs has been demonstrated using the testing techniques that are to be employed.
>
> Third, make sure that the scoring of response relates directly to what is being tested.
>
> Finally, do everything possible to make the test reliable. If a test is not reliable, it cannot be valid. (Hughes 2003: 33-34)

With great efforts, the validity of test can be moved on to a higher level of standard. Therefore, the test will be more useful for both candidates and examiners.

The Relationship between Reliability and Validity

Reliability and validity are critical for tests, and are sometimes referred to as essential measurement qualities since the primary purpose of a language test is to provide a measure that can be interpreted as an indicator of an individual's language ability. They are two closely related

ideas and researchers have many theories about their relationship. Hughes (2003: 50) and Bachman and Palmer (1996: 23) suggest that, in order to be valid, a test must provide consistently accurate measurements, which means reliability is a necessary condition for validity, and hence for usefulness. However, reliability is not a sufficient condition for validity. In other words, a reliable test may not be valid at all. For some other researchers, test validity is requisite to test reliability. If a test is not valid, then reliability is meaningless (OPTISM n.d.). That means if a test is not valid, there is no point in discussing reliability since test validity is required before reliability can be considered in any meaningful way. It is the same the other way round. If a test is not reliable, it is also not valid (OPTISM n.d.). Figure 3 can explain the relationship between reliability and validity clearly:

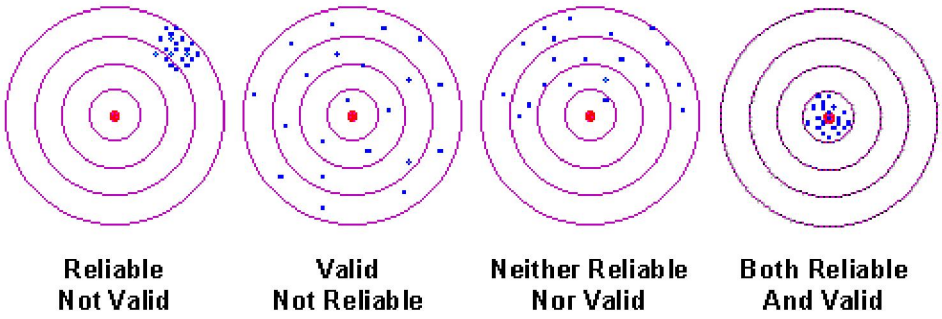

Figure 3: The Relationship between Reliability and Validity (Research Methods Knowledge Base 2006)

The center of the target is the concept that examiners are trying to measure. Every shot represent one candidate that is to be measured. If you hit the center of the target, you measure the concept perfectly for a candidate. Otherwise, you do not. The more you are off for that person, the further you are from the centre. In Situation 1, you are consistently hitting the target, but off the center of the target. That means you are consistently measuring the wrong value for all respondents. In this case, the measure is reliable, but it is not valid. In Situation 2, you are randomly hitting the target, so the hits are spread disorderly. You seldom hit the center of the target, but on average, you are getting the right answer for the group. Under this circumstance, the test is valid but not consistent. Situation 3 shows that the hits are not randomly spread. Moreover, you consistently miss the center. The measure in this case is neither reliable nor valid. In the last situation, you consistently hit the center of target. In this case, the measure is both reliable and valid.

In order to exert the usefulness of a test, both reliability and validity are essential parts to which test designers should make great efforts.

2.2.3 Authenticity

When making inferences about test takers' language ability, the inferences are supposed to generalize to those specific domains in which the test takers are likely to need to use language, in other words, in a target language use domain. Bachman and Palmer define a *target language use (TLU) domain* as "a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize" (1996: 44). The TLU domain is an essential element of the usefulness of a test.

In order to justify the usefulness of language tests, it is important to demonstrate that performance on language tests corresponds to language use in specific domains other than the test itself. Authenticity is just to measure the extent of one aspect of demonstrating, the correspondence between the characteristics of TLU tasks and those of the test task. Authenticity is defined as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (Bachman & Palmer 1996: 23). A TLU task here refers to an activity that individuals are involved in, using the target language for achieving a particular goal or objective in a particular situation. A more vivid explanation of authenticity is shown in Figure 4.

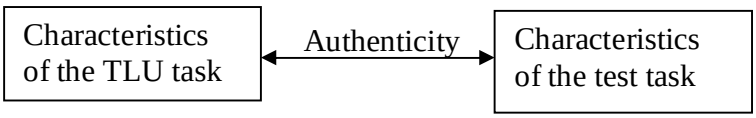

Figure 4: Authenticity (Bachman & Palmer 1996: 23)

For example, if a test examines communicative ability, authenticity here refers to the degree of correspondence of the characteristics of the test task to the features of the communication task. If the given test construct closely resembles the situation a test-taker would face in the TLU domain, the test is more authentic. In other words, the test task is likely to be enacted in the real world.

In a test, evidence of authenticity may be presented in the following ways:

- The language in the test is as natural as possible.

- Items are contextualized rather than isolated.
- Topics are meaningful (relevant, interesting) for the learner.
- Some thematic organization to items is provided, such as through a story line or episode.
- Tasks represent, or closely approximate, real-world tasks. (Brown 2004: 28)

(1) The language used in the test should be as natural as possible because test takers read instructions and items through the language, whether it is their first language or target language. Test designers should avoid using language with academic or technical terms. (2) In the real world, we seldom use single items; rather, we use items in phrases or sentences. Thus, items in the test should be contextualized rather than isolated. (3) Authentic means the test task is also used in the daily learning process, which means the topics of the test should be closely related to test takers' daily life. The topics should be meaningful, relevant, or interesting for the test takers. (4) In test tasks like cloze, there are always paragraphs of a story and test takers are required to fill in blanks. However, if some thematic organization to items is not provided, test takers would have no idea about what the plot of the story is, and they would not be able to finish the task. Therefore, thematic organization of items should be provided in the test. (5) This evidence is the most important for authenticity in a test that the test tasks should be close to real-world tasks.

In attempting to design a test task with authenticity, the test designer should first identify the critical features that define tasks in the TLU domain. This recognition serves as a framework for the task characteristics. Test tasks that have these critical features are then designed and selected.

A language test is said to be authentic when it mirrors as exactly as possible the real life non-test language tasks. Testing authenticity can be divided into 3 categories, which are input (material) authenticity, task authenticity, and layout authenticity. Input authenticity means that authenticity should be present in the test material, and it further falls into three aspects, situation authenticity, content authenticity, and language authenticity. Task authenticity forms the cornerstone of test authenticity. In authentic tasks, the emphasis should primarily be on the proficiency levels of the population. The layout of the test paper should also be authentic. According to Bo (2007: 5), the

most usual way to make authentic layout of test paper is by presenting pictures. Vivid pictures can be used to test those productive skills as speaking and writing.

2.2.4 Interactiveness

Interactiveness is another important element in the quality of usefulness. Bachman and Palmer (1996: 25) define interactiveness as "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task". Individual characteristics, such as test taker's language ability (language knowledge and strategic competence[1], or metacognitive strategies), topical knowledge, and affective schemata[2], are most relevant for language testing. These can be shown as in Figure 5.

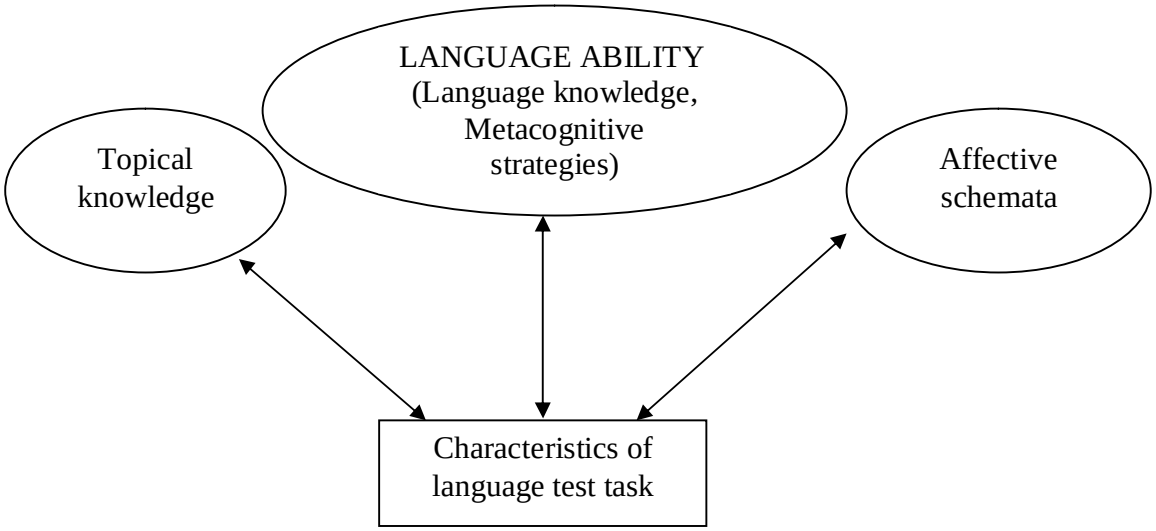

Figure 5: Interactiveness (Bachman & Palmer 1996: 26)

There are interactions between language ability, topical knowledge and affective schemata, and the characteristics of language test task. Authenticity refers to the characteristics of test tasks and features of TLU tasks, while interactiveness refers to the interaction between the test taker and the test task. Many types of test tasks may involve the test taker in a high level of interaction with

[1] Bachman and Palmer (1996: 70) conceive strategic competence as "a set of metacognitive components, or strategies, which can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities". In other words, strategic competence, or metacognitive components provides an essential basis for designing and developing test tasks and for evaluating the interactiveness of the test tasks.

[2] Affective schemata can be considered as the affective or emotional correlates of topical knowledge (Bachman & Palmer 1996: 65). Students' affective schemata can influence their performance on tasks when they deal with emotionally charged topics, such as abortion, gun control, or national sovereignty.

the test input, such as responding to visual, non-verbal information. However, test taker's language ability cannot be defined based on his performance in the test unless this interaction requires the use of language knowledge. Therefore, interactiveness is a critical quality of language test tasks since it is closely related to construct validity.

The Common Ground of Authenticity and Interactiveness and Their Relationship with Construct Validity

According to Bachman and Palmer(1996: 28-29), there are some points that authenticity and interactiveness share with each other in designing, developing, and using language tests. Firstly, since both authenticity and interactiveness measure extent and degree, we can just say relatively more or relatively less authentic or interactive, rather than authentic and inauthentic, or interactive and non-interactive. Secondly, when we talk about authenticity and interactiveness, we must consider three aspects of characteristics: characteristics of the test takers, characteristics of the TLU task, and characteristics of the test task. Thirdly, certain test tasks are relatively useful for their purpose even with low authenticity or interactiveness. Fourthly, our understanding of a test task's authenticity and interactiveness is just a guess since different test takers have different characteristics that they perform differently in the same test. Fifthly, the lowest acceptable levels that we specify for authenticity and interactiveness depend on the specific testing situation, and they must be balanced with those for the other test qualities.

As Bachman and Palmer (1996: 29) suggest, "[a]uthenticity, interactiveness, and construct validity all depend upon how we define the construct 'language ability' for a given test situation". Authenticity is about the correspondence of test task and TLU task, so it is of course closely related to content validity. Moreover, authenticity provides a means for investigating the extent to which score interpretations generalize beyond performance on the test. Since investigating the generalizability of score interpretations is an important part of construct validity, authenticity and construct validity are linked. According to Figure 2, both interactiveness and construct validity have something to do with language ability, which includes language knowledge, strategic competence, metacognitive strategies, and furthermore, the topical knowledge. The degree of how interactiveness corresponds to construct validity depends on how we define the construct and on the characteristics of the test takers.

2.2.5 Impact

Another quality of tests is their impact on society and educational systems as well as upon the individuals within those systems. A test is to serve a specific purpose, thus the test scores also imply values and goals, and they have consequences. As Bachman (1990: 279) points out, “tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or of society at large”. Thus, whenever we use tests, our choices have specific impact on both the individuals and the system involved.

There are two levels of the impact of test use. At a micro level, it is the individuals that are affected by the particular test use. At a macro level, the educational system and the society are affected by the particular test use.

Washback

When we deal with the impact of tests, one aspect should be mentioned first. Bachman and Palmer name it as “washback” (1996: 30) while Hughes calls it as “backwash” (2003: 1). This concept pertains to the effect of testing on teaching and learning, and it can be maleficial or beneficial. If in a test, the test designer asks candidates to write a composition to test their oral ability, the test would bring maleficial washback. It is because writing a composition is actually testing writing ability, although oral ability is a comprehensive ability that may also include writing ability, it will give the candidates the impression that the skill of speaking can be ignored in the classroom learning. Therefore, the test itself also cannot reach its purpose. “Cram” courses and “teaching to the test” are examples that maleficial washback bring to the classroom (Brown 2004: 29). On the other hand, beneficial washback or positive washback “depends in part upon factors such as the importance of the test, the status of the language being tested, and the purpose and format of the test” (Weigle 2002: 54). The test itself cannot ensure beneficial washback in consideration of the possibility that many factors outside the test may affect washback, like teacher’s personal beliefs, institutional requirements, and student expectations.

Impact on Test Takers

Test takers are among those individuals who are most directly affected by test use. According to Bachman and Palmer (1996: 31), mainly three aspects of the testing procedure affect test takers:

· the experience of taking and, in some cases, of preparing for the test,

· the feedback they receive about their performance on the test, and

· the decisions that may be made about them on the basis of their test scores.
(Bachman & Palmer 1996: 31)

Firstly, the experiences of preparing for and taking the test have the potential possibility of affecting characteristics of test takers including personal characteristics, the topical knowledge, affective schemata, and their language ability. In high-stakes tests such as national examinations or standardized tests, test takers may spend several weeks or even months preparing for the test. Some high-stakes nation-wide public examinations are used as placement tests or proficiency tests, just as the test being analyzed in this essay, which is used for selection and labelling of different test takers within different levels. In these examinations, teaching may be focused on the syllabus of the test for up to several years before the actual test, and the techniques in the test will be practiced in class. Moreover, the experience of taking the test also has impact on test takers. The test taker's perception of the TLU domain, areas of language knowledge, and use of strategies may be affected by the test.

Secondly, the feedback that test takers receive about their performance in the test is likely to affect them directly. Therefore, feedback should be as relevant, complete, and meaningful to the test taker as possible. In most situations, the feedback of test performance is a score. However, in order to make beneficial impact on test takers, rich verbal description of the score, the actual test tasks, and the test taker's performance are also needed.

Finally, the decisions that may be made about the test takers based on their test scores may directly affect them in various ways. In low-stakes tests, the result of test will help students to discover their areas of strength and weakness so they know what more has to be done. On the other hand, in the examinations like University Entrance Examination, the result will directly determine whether a student can be admitted to a university or not. Some proficiency tests related to job hunting will also determine whether one can be employed or not. All these decisions have serious consequences for test takers. Therefore, fair decisions, which are with equally appropriate, regardless of individual test takers' group membership, should be made. Fair test use also pertains to the relevance and appropriateness of the test score to the decision, as well as whether

and by what means test takers are fully informed about how the decision will be made and whether decisions are actually made in the way described to them.

Impact on Teachers

Test users are the second group of individuals who are directly affected by tests, including test designers, test examiners, and administrators. In an instructional program, the test users that are most directly affected by test use are teachers. Impact on the program of instruction is considered as washback for test users. Most teachers are familiar with testing influence on their instruction. The term 'teaching to the test' is unavoidable for most situations for teachers. It implies "doing something in teaching that may not be compatible with teachers' own values and goals, or with the values and goals of the instructional program" (Bachman & Palmer 1996: 33). If teachers feel that what they teach is not relevant to the test, it must be an instance of low-test authenticity, in which the test has maleficial washback on instruction. Therefore, a useful test should be provided to minimize the potential for negative impact on instruction.

Impact on Society and Education Systems

In addition to test takers and test users, the society and education systems are also influenced by the impact of tests. In a second or foreign language testing, the consideration of values and goals is especially complex since the values and goals that inform test use may vary from one culture to another. Different cultures have different aspects to be valued. According to Shohamy (1998: 332), some tests reflect the social condition while some other tests reflect the political condition. Values and goals also change over time. Secrecy and access to information, privacy and confidentiality were never considered once, but they are now considered as basic rights of test takers.

High-stakes tests, which are used to make major decisions about large numbers of individuals, are particularly likely to have consequences not only for the individual stake holders, but also for the educational system and society. An achievement test may have potential impact on the language teaching practice and language programs. A test with intended purpose may also have impact on the society. As Shohamy (1998: 332) argues, "[…] the act of testing is not neutral. Rather, it is both a product and an agent of cultural, social, political, educational and ideological agendas that shape the lives of individual participants, teachers and learners". Tests like TOEFL

and IELTS are used to screening students applying for studying in English speaking countries as well as individuals applying for immigration. In this way, language tests have wider social and political implications as well.

2.2.6 Practicality

Practicality is very different from the qualities mentioned above which pertains to the ways in which the test will be implemented and whether it will be developed and used at all, rather than the uses that are made of test scores. Bachman and Palmer (1996: 36) define practicality as "the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities". This relationship can be represented as in Figure 6.

$$\text{Practicality} = \frac{\text{Available resources}}{\text{Required resources}}$$

If practicality ≥ 1, the test development and use is practical.
If practicality ≤ 1, the test development and use is not practical.

Figure 6: Practicality (Bachman & Palmer, 1996: 36)

For any given situation, if the required resources for implementing the test exceed the available resources, the test will be impractical. The test designer should either reduce required resources or increase available resources to make the test more practical. In reverse, the test is practical.

There are three general types of resources to assess practicality, human resources, material resources, and time. Human resources include test writers, scorers or raters, test administrators, and clerical support. Material resources refer to space (for example, rooms for test development and test administration), equipment (for example, typewriters, word processors, tape and video recorders, computers), and material (for example, paper, pictures, library resources). Time is composed of development time (time from the beginning of the test development process to the reporting of scores from the first operational administration) and time for specific tasks (such as designing, writing, administering, scoring, analyzing). The specific resources required will vary

from one situation to another, thus, practicality can only be determined for a specific testing situation.

Brown (2004: 19) defines a practical test as one that "is not excessively expensive, stays within appropriate time constraints, is relatively easy to administer, and has a scoring/evaluation procedure that is specific and time-efficient". High-stakes tests require a great deal of resources, so they are considered costly and time-consuming. However, it is impossible to spend too much money on a test, or ask test takers to spend hours on finishing the test, or require examiners to take hours to evaluate a test paper, or only computer can score the test while the test is taken far away from the nearest computer. Therefore, it is not surprising that "some test users may search for ways to avoid less practical tests if they believe other tests can serve the same purpose" (Gennaro 2006: 2). In other words, the value and quality of a test sometimes depends on quite detailed, practical considerations.

3. Analysis and Discussion

The analysis mainly consists of two parts, the system of the test and its impact, and the test usefulness. There are three aspects to be discussed within the system of the test, which are the organization of the test, the implementation of the test, and the reform of the test. In the reform part, there is a comparison between the old TEM-8 and the new TEM-8. The test usefulness part pertains to reliability, validity, authenticity, interactiveness, impact and practicality. Since this essay emphasizes the impact, the impact is analyzed and discussed as a separate part.

3.1 The System of the Test and Its Impact

When referring to the system of a test, we are talking about who makes the test, who organizes the test, how the test is organized, what is the purpose of the test, what is the content of the test, how the test is scored, and some other important issues when we deal with testing. In this essay, the system of the test is divided into three parts, the organization of the test, the implementation of the test, and the reform of the test. The impact of the system on students is discussed as well when analyzing the system.

3.1.1 The Organization of the Test

As mentioned in the Introduction, TEM-8 was set up by the State Education Commission in 1991, and has been organized by the Higher Education Institution Foreign Language Major Teaching Supervisory Committee since then. Every year, the English group members of the supervisory committee design the test and then send the test to colleges and universities all around the country.

However, the Higher Education Institution Foreign Language Major Teaching Supervisory Committee has no official web site. They just publish *The Syllabus of Test for English Majors Band 8* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2005), and inform of the time of testing in that year. There is no official way to make the test content public, only after scoring, some examiners will take pictures of the test and put them on internet. There are no official answers to the test as well, and this is one of the aspects that students complain about, since they have no idea where they have made mistakes. One of the students who took questionnaires says, “there are so many questions we are not sure of the answers”. Fortunately, English experts will do the test themselves and publish their answers on internet, and generally speaking, different experts have the same answers to objective parts of the test like multiple choices. For the more subjective parts, there are some slight differences, but as a whole, they are similar to each other, at least with the standard or scoring. Both teachers and students accept the experts’ answers.

3.1.2 The Implementation of the Test

Every year, after the English group members of the supervisory committee design the test, the test papers are sent to colleges and universities all around the country. Every higher education institution is in charge of the test process within their own colleges or universities. They arrange the classrooms and supervisors, check up the identifications of the test takers, supervise the using of equipments in the classrooms, distribute as well as collect test papers, and send the test papers back to the supervisory committee.

The Participants of the Test

The main participants of TEM-8 are Grade Four English major students of higher education institutions that are confirmed by the Ministry of Education. In the meantime, those non-English major students who have passed CET-6 (College English Test Band 6) can also attend the TEM-8. However, this number of students is very small. CET is one of the most pervasive English tests in China, which is set for all college students, no matter they are English major or not. Therefore, CET-4 and CET-6 are relatively much easier for English majors. The difficulty of CET-6 is probably equal to TEM-4, thus TEM-8 is much more difficult than CET-6. This is why there are so few non-English major students taking TEM-8. For all students who have taken TEM-8, they have one chance of taking the re-sit examination. Those who fail in the first year can attend the next year's TEM-8 testing.

The Time of the Test

The time of the test brings much inconvenience to the students. 60%[3] of the English majors that have answered my questionnaire state that the test taken in Semester 8 is not suitable. Although the test means to measure the implementation of the syllabus after the four years of teaching, most of the 60% English majors indicate that it is better to set the test in Semester 7. There are two reasons for this indication. Firstly, during the fourth year of college or university, students are preparing for the examination for further education, working on their graduation paper, and hunting for jobs, especially in Semester 8 when the Spring Festival is over. They have so many things to do in the last year and semester. With Question 15 in my questionnaire, 14% of the English major students show that the test influences their preparation for the graduation paper, and 30% of them indicate that the test influences their hunting for a job. This actually affects their ordinary learning and life. Secondly, the result of the test comes out in late May or early June, and only by then, students who have passed the test can get their certifications. However, many jobs need this certification when they are applied for. Therefore, students may lose many chances of good jobs. One more thing is that, if a student fails the test, it means that if he really wants to pass the test, he has to take the next year's test again, but by that time, he may be

[3] The data of questionnaire is set in Appendix 1. Numbers in percentage in the analysis usually refer to the percentage of students who have taken my questionnaire, unless I mention other groups specifically.

working already, studying abroad or just be far away from the college. It is inconvenient or even unpractical to attend the re-sit examination.

The Test Framework

The test contains six parts, i.e. Listening Comprehension, Reading Comprehension, General Knowledge, Proofreading and Error Correction, Translation, and Writing. Among these six parts, Listening Comprehension and Translation have subsections. Various testing techniques are adopted, such as multiple choice questions, and gap filling. As far as score is concerned, the Listening Comprehension, Reading Comprehension, Translation, and Writing take up to 20% of the total score respectively, while General Knowledge and Proofreading and Error Correction account for 10% respectively (see Table 1).

Table 1: The framework of TEM-8

part	Test item		Format	Number of items	Percentage of scoring	Time (min.)
		Mini-lecture	Gap Filling	10	10%	
I	Listening Comprehension	Interview	Multiple Choice	5	5%	35
		News Broadcast	Multiple Choice	5	5%	
II	Reading Comprehension		Multiple Choice	20	20%	30
III	General Knowledge		Multiple Choice	10	10%	10
IV	Proofreading and Error Correction		Gap Filling	10	10%	15
V	Translation	Chinese to English	Translation	1	10%	60
		English to Chinese	Translation	1	10%	
VI	Writing		Passage Writing	1	20%	45
Total				63	100%	195

In terms of scoring, the objective parts, including Interview, News Broadcast, Reading Comprehension, and General Knowledge, take up 40% score of the whole test, while subjective parts, such as Mini-Lecture, Proofreading and Error Correction, Translation, and Writing take up 60% score of the whole test. Therefore, there is a lot of writing the students should do. It is obvious that the students have to do more thinking than in a test with 90% objective items.

The test begins at 8:15 a.m. in the early morning, and lasts for more than 3 hours. Although there are so many aspects to be tested, the long time of intensive work makes students exhausted. It is hard to imagine what the scene will be if an oral part is included in the test as well in the future.

There is another thing that should be mentioned here. TEM-8 is not like other ordinary tests where the test papers are handed out at the beginning of the test and collected all together at the end of the test. In contrast, the test papers are handed out and collected in many steps:

(a) The test papers of section I (without the Mini-Lecture part), II, III, IV and the answer sheets are handed out.
(b) After listening to the Mini-Lecture, the test papers of this part are handed out but answer sheet 1 for this part are collected after 10 minutes.
(c) After the time for the section III (General Knowledge) has run out, in other words, after 75 minutes, the answer cards for these four parts are collected.
(d) After the answer cards are collected, for each section left, the test papers and answer sheets for that part are handed out at the beginning of that part and are collected at the end of that part.

In a word, the test papers and answer sheets are handed out and collected five times. Apparently, the supervisory committee sets the test process like this to make sure there is less possibility for students to cheat, and also to control the students' time for each section. However, this complex process brings much trouble for students.

Firstly, they may feel stressed because of these procedures. In an ordinary test, students get all the test papers and answer sheets at the beginning of the test, and hand them in together at the end of the test. Therefore, they can arrange the time for different sections in their own way. For example,

in this test, if without these procedures, students who use less time of General Knowledge can spend more time on Reading Comprehension. However, in this test, students have to answer different sections in respectively precise times. Therefore, they feel different towards different sections (see Figure 7).

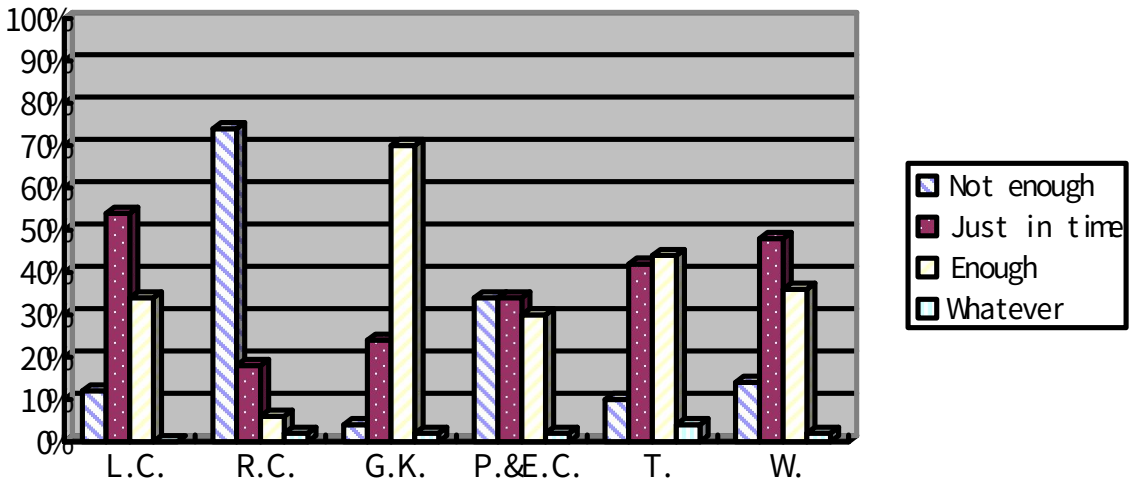

Figure 7: The Feeling of Time limitation in Different Sections
(L.C.—Listening Comprehension; R.C.—Reading Comprehension; G.K.—General Knowledge; P. & E.C.—Proofreading and Error Correction; T.—Translation; W.—Writing)

For Listening Comprehension, 54% of the students think they were just in time. For Reading Comprehension, 74% of the students think they did not have enough time for this part. For General Knowledge, 70% of the students think they had enough time for this part. For both Proofreading and Error Correction, 34% of the students think they did not have enough time or were just in time to finish this section. For Translation, 42% of the students feel they were just in time to finish this section while 44% of them feel they had enough time for this section. For the last section, Writing, 48% of the students feel they were just in time for this section. Overall, students feel stressed about the limitation of the time for each section. On the other hand, since the supervisors have to hand out, and collect test papers and answer sheets, they have to walk around the classroom, and this brings stress to students as well. The atmosphere can be very intense, and that is one of the reasons why some students feel extremely stressed in the classrooms when taking the test.

Secondly, most students perform better without being disturbed. However, in this test, they are disturbed many times. The fluency of thinking is cut off, and this may have a great influence on the performance of students, thus to affect their scores of the test.

The Requirement of the Test

In *The Syllabus of Test for English Majors Band 8* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2005: 2-4), there is a list of requirements for each section of the test. The requirements are very specific and detailed. For each section, the syllabus brings the level of knowledge students should have acquired, as well as the level of language using ability (see Appendix 3).

Listening Comprehension

From the requirement of the listening part, we can see that the syllabus has a high standard for English major students with their listening ability. Although it means to test students' listening ability, it requires knowledge of all aspects as well, such as politics, economy, history, culture and education, and so on. Since the syllabus particularly points out foreign media, such as VOA, BBC, and CNN, material from these media have become an important resource for listening class in general teaching.

Listening Comprehension consists of three parts: mini-lecture, interviews, and news broadcast. Among these three tasks, the mini-lecture is considered to be the most difficult task. 12% of the students who took the questionnaire directly point out that this part is very difficult. Students just get a piece of paper with no characters on it before the recorder runs. They have to write down what they have heard as soon as possible and as much as possible. They can just hear the lecture once, and after that, they will get a test paper and do gap filling. However, not the whole lecture is on the paper, just the brief version. 10 blanks are given to be filled. Many students, including me, are mind blank after listening to the lecture since we have no idea about the topic of the lecture before listening to it, and in addition, the speed of the lecture is almost the same as in a lecture in an English-speaking country. The 10 blanks are supposed to be related with the main clue of the lecture. However, the lecture is approximately 900 words long, thus it is difficult for students to write down the main clue when they hear the passages for the first time.

However, in spite of its level of difficulty, 80% of the students think that the Listening Comprehension section can really reflect their listening ability.

Reading Comprehension

As mentioned above, 74% of the students feel that they did not have enough time for the Reading Comprehension. Time limitation is of course a reason. There are about 700 words in each text (four texts), so considering the time limit, students are supposed to read the content with the speed of around 150 words one minute. In addition, they have to think about the answers as well. Sometimes they cannot directly give the answer but hesitate and think longer. Furthermore, without enough time is not only because of the time limitation, but also because of the level of difficulty of this section. According to the requirement, candidates not only need to understand the general idea of the article, but also need to be able to analyze details. Referring to the content of the articles, they are in large scale. Politics, economy, history, culture, education, science are all included. In Question 3, about the difficulty of the whole test, 10% of the students directly indicate that Reading Comprehension is difficult and needs a large scale of background knowledge. In addition, in Question 8, about whether they are familiar with the topics in the test, 48% of the students chose no. Furthermore, since this section is to test students' reading ability, a great amount of vocabulary is required. Obviously, students with larger vocabulary can read the texts more quickly and understand the content much better. Therefore, in order to perform better in this section, students should do much practice at other times.

General Knowledge

This section is the only section that candidates can prepare for with particular material rather than just practicing for training the abilities. Since the content of this section is described in the requirement, all what students should do is to look for all the material and remember them in mind. However, referring to the material, since the scale of the content is still very large, students complain that there are too many things to be remembered, and it is very easy to get confused with similar countries, or literary works. Therefore, just 10% of the students think that this section can really reflect their language ability. Nevertheless, the general knowledge is still a very important part of language knowledge that students should acquire.

Proofreading and Error Correction

The short paragraph consists of about 250 words, and within each line the test paper indicated, there is an error. Students are required to correct the errors by deleting a word, changing a word, or adding a word. Although there are 15 minutes for 10 items, students still think they need more time. 34% of the students feel that there was not enough time to for them to finish this section, and another 34% of the students express that they were just in time to finish this section. This is mainly because of the difficulty of this section rather than just the time limitation. In this section, the students are tested on a great amount of linguistic knowledge, for example, the structure of the sentences and paragraph, the vocabulary, and the lexical chunks. Students are required to be good at linguistics, be able to connect items within the context, and be sensitive to errors. This section is also closely related to students' reading ability.

Translation

In this section, 42% of the students indicate that they were just in time to finish this part, and 44% of the students feel that there was enough time for them to finish this section. Although the time limitation is not so strict, the level of difficulty of this section is not lower than that of other sections. In this section, students need a great amount of background knowledge, such as politics, economy, history, culture and science. Furthermore, knowing the background knowledge is not enough, the students should be familiar with the English expression or Chinese expression for the knowledge. This is especially important for some titles of events, departments, and meetings. Without these expressions, the translation process will not be smooth. This section is to test candidates' reading ability as well as writing ability. Students are supposed to be very familiar with both Chinese language and English language knowledge, know the similarities and differences between the two languages well, and be capable of translating either language to the other.

Writing

24% of the students express that they were just in time to finish this part. They have to follow the instruction and take time to think about how to arrange the essay. If they arrange the time properly, they should be able to finish in time. In this section, the purpose is to assess students' writing ability. However, writing ability is a comprehensive ability, including abilities of reading,

understanding, and writing. Students should have certain background knowledge[4], imagination, and logic. This section can reflect students' level of acquired language knowledge, their ability to understand the knowledge, as well as their ability to use the knowledge. Therefore, it is an integrative test. 35% of the students point out that they think this section can reflect their language ability.

The Score of the Test

The scores of the test are divided into four categories. The total score of the test is 100, and the ones below 60 are failed. The four categories are distinguished as in Table 2.

Table 2: The Four Categories of the Score of the Test

Scores	Labels
80-100[5]	Distinction
70-79	Merit
60-69	Pass
0[6]-69	Fail

Every year, the national pass rate of TEM-8 is between 40% and 50%. Some better colleges and universities can reach above 80%. Therefore, these colleges take the high pass rate as their advantage when they enrol freshmen after the National College Entrance Examination.

[4] The 'background knowledge' here refers to schematic knowledge, which includes relative knowledge, memory and experience, and provides relative material when we interpret new information.

[5] Actually, the score of 100 is an ideal situation, since there are subjective parts in this test, even the students are all correct with the objective parts, there is little possibility that they can be perfect at subjective parts as well. In addition, the marking of the subjective parts stops students getting full score too. In Translation and Writing, the students are required to use appropriate choice of words, and have a variety in sentence patterns. The structure should be logically organized and relatively few significant errors of vocabulary, spelling, and punctuation should be found (Enfamily 2006). Therefore, and in fact, no one ever has the full score.

[6] The score of 0 is also an ideal situation as in footnote 3. There is probably only one possibility that the students can get 0 score. That is when the student misses the test.

The result of the test comes out in late May or early June. However, as mentioned above, unlike CET-4 and 6, TEM-8 has no official web site. Students cannot track their scores on line. Every year, after the Higher Education Institution Foreign Language Teaching Supervisory Committee English Group finish scoring, the results will be mailed to every college and university as well as the certifications. Every college and university is in charge of informing students of their results and gives them the certifications.

3.1.3 The Reform of the Test

There are some changes between the two syllabuses, mainly focusing on the content of the test (see Table 3). As a whole, there are four main changes between the old syllabus and the new syllabus.

Table 3: The Changes between the Old Syllabus and the New Syllabus

Sections	The old syllabus	The new syllabus
Listening Comprehension	**Section A** Talk (5 items) **Section B** Conversation (5 items) **Section C** News Broadcast (5 items) **Section D** Note-taking & Gap-Filling (10 items) Time: 40 minutes Percentage of score: 25%	**Section A** Mini-lecture (10 items) **Section B** Conversation (5 items) **Section C** News Broadcast (5 items) Time: 35 minutes Percentage of score: 20%
Reading Comprehension	**Section A** Careful Reading (15 items) **Section B** Speed Reading (10 items) Time: 40 minutes Percentage of score: 25%	Reading Comprehension (20 items) Time: 30 minutes Percentage of score: 20%
General knowledge	Not included	10 items about English knowledge Time: 10 minutes Percentage of score: 10%
Proofreading and Error Correction	10 language disorders Time: 15 minutes Percentage of score: 10%	The same
Translation	**Section A** Chinese to English **Section B** English to Chinese Time: 60 minutes Percentage of score: 20%	The same
Writing	Controlled composition writing (300 words) Time: 60 minutes Percentage of score: 20%	Controlled composition writing (400 words) Time: 45 minutes Percentage of score: 20%

Sections	The old syllabus	The new syllabus
Percentage of objective and subjective scores	Objective: 40% Subjective: 60%	Objective: 40% Subjective: 60%
Time for test	215 minutes (divided into morning section and afternoon section)	195 minutes (just morning section)
The order of the sections	Listening Comprehension → Proofreading & Error Correction→ Reading Comprehension → Translation → Writing	Listening Comprehension→ Reading Comprehension→ General Knowledge→ Proofreading & Error Correction→ Translation→ Writing

(a) The quantity of vocabulary that students should have acquired in the old syllabus is above 10,000 words, while that in the new syllabus is above 13000 words, which means the requirement for students has risen to a higher standard. Students have to work harder in and after classes.

(b) The total time for the test is reduced. In the old syllabus, the total time is 215 minutes, and the test is divided into two parts, the morning part and the afternoon part. In the morning part, students are tested with Listening Comprehension, Proofreading and Error Correction, and Reading Comprehension sections, and in the afternoon, they are tested with Translation and Writing. However, in the new syllabus, the total time is reduced to 195 minutes, and all sections are supposed to be finished together in the morning. The total number of the items is not changed, which means in the new syllabus, students have to finish the same number of items within less 20 minutes' time. The difficulty level of the test has risen to a higher stage.

(c) The new syllabus has more requirements for the sections of Listening Comprehension, Reading Comprehension, and Writing. For Listening Comprehension, in the old syllabus, the proportion of selected response (multiple choice) is larger than that of limited response (gap filling), while in the new syllabus, the two proportions are the same, which means students have to commit themselves harder to the test, and answer with more or larger scale of language knowledge. For Reading Comprehension, in the old syllabus, Careful Reading tests students' ability of reading carefully, and Speed Reading tests students' ability to read fast, while in the new syllabus, there are just comprehensive tests. The purpose of this is to reflect more truthfully

the students' psychology of reading and encourage students using various ways to read according to different needs in the process of reading. For Writing, in the old syllabus, students have to write 300 words in 60 minutes, while in the new syllabus students have to write 400 words in just 45 minutes. Time is reduced while the quantity is enlarged. This requires students to acquire better writing ability to deal with this section.

(d) The new syllabus adds a new section, General Knowledge. According to the *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000: 16), English major should have three aspects of courses, English major skill courses, English major knowledge courses, and relative majors knowledge courses. Among them, English major knowledge courses refer to courses which are related to English language, literature and culture, such as English Linguistics, English Lexicology, English and American literature, English-speaking Countries' Society and Culture. General Knowledge is added to test this aspect of knowledge. Therefore, the test can assess English major senior students' professional qualities in a more comprehensive way. For students, this means they have more things to prepare. They need to pay attention to relative knowledge during their daily learning and keep the knowledge in mind.

As a whole, compared with the old syllabus, the new syllabus brings forward more and higher requirements for the students. In order to pass the test, students should work harder on all aspects of English knowledge, and of course, they have to do more practice to strengthen their ability so as to get a better score. Conversely, with higher requirement, the professional level of the students will also be higher.

3.2 The Test Usefulness

After the discussion and analysis of the system of TEM-8, this essay uses more theoretical ways to analyze and discuss the usefulness of TEM-8 in order to find out the impact of the test on both micro level and macro level.

Before its usefulness, the purpose of TEM-8 should be discussed. As mentioned above, TEM-8 is set up to assess and evaluate the actual performance of *Higher Education Institution English*

Major English Teaching Syllabus (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000) for senior English major students. In this way, the test should be considered as an achievement test. However, the test sets a certain standard to measure students' language ability as well. The content of the test has little to do with the content of classroom teaching, as the test measures abilities and methods rather than the content. Furthermore, for those English majors who get the certification of TEM-8, it is easier to get the jobs requiring the certification. In this way, the test should be considered as a proficiency test.

3.2.1 Reliability

As Bachman and Palmer (1996: 135) indicated, "probably the most important consideration in setting a minimum acceptable level of reliability is the purposes for which the test is intended". Therefore, as a relatively high-stakes test, the test designers of TEM-8 mean to set the minimum acceptable level of reliability very high. To measure the degree of reliability in TEM-8, two components of test reliability should be paid attention to, the performance of candidates from occasion to occasion, and the reliability of the scoring.

For the old TEM-8 syllabus, according to Lan and Huang's (2004: 2) data, from 1997 to 2003, the national pass rates are respectively 55.6%, 53.6%, 54.2%, 55.7%, 58.2%, 59.6%, and 51.8%[7]. The rates are in a certain range. For the new TEM-8 syllabus, the national pass rate is relatively consistent as well. The Figure 8 shows the national pass rate of TEM-8 in the latest five years (from 2005[8] to 2009).

[7] The enrolment expansion of college and university students began in 1999, therefore, in 2003, the first batch students were in Grade 4 and attended the test. The larger number of students and their different levels of language ability influenced the national pass rate.

[8] The national pass rate in 2005 is much higher than other years. Experts explain that since 2005 was the first year the new syllabus for TEM-8 taken into action, students had to adapt themselves to the new item types, therefore, the level of difficulty was a little lower than predetermined. However, in the following years, the level of difficulty has been raised to the standard level, therefore, the national pass rate dropped a little. On the other hand, the enrolment expansion of college and university students since 1999 also affects the result since different students have different levels of language ability.

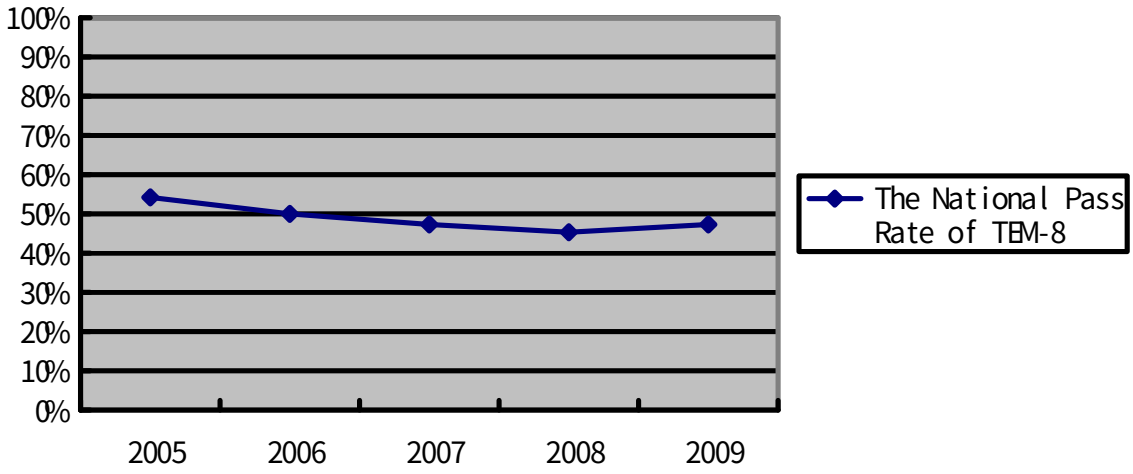

Figure 8: The National Pass Rate of TEM-8

The rate dropped a little in 2006, 2007, and 2008, however, in 2009, the rate rose up. These five years' rates are around 45% to 50%. It seems that candidates' performance in different years is relatively consistent, at least in latest five years. As a whole, the test is relatively high stability consistent, in other words, consistent over time. Both with the old syllabus and with the new syllabus, the reliability estimates are well within the desirable range and substantial.

There are two more considerations that need to be kept in mind when setting high minimum acceptable level of reliability – "the way the construct has been defined and the nature of the test tasks" (Bachman & Palmer 1996: 135). That means only when the test focuses on a relatively narrow range of language abilities with relatively uniform test tasks, could the test achieve higher levels of reliability. In the Writing section of the TEM-8 test, a controlled composition is adopted (see Figure 9, the Writing section of the test paper of TEM-8 in 2009)

Figure 9: 2009 TEM-8 Test Paper Writing Section (*2009 TEM-8 Test Paper* 2009)

Mandarin, or Putonghua, is the standard service sector language in our country. But recently, employees at a big city's subway station have been busy learning dialects of other parts of the country. Proponents say that using dialects in the subway is a way to provide better service. But opponents think that encouraging the use of dialects in public counters the national policy to promote Putonghua, what is your opinion? Write an essay of about 400 words on the following topic: Are Dialects Just as Acceptable in Public Places?

In the first part of your essay you should state clearly your main argument, and in the second part you should support your argument with appropriate details. In the last part you should bring what you have written to a natural conclusion or make a summary.

Marks will be awarded for content, organization, grammar and appropriateness. Failure to follow the above instructions may result in a loss of marks.

Write you essay on ANSWER SHEET FOUR.

As shown in Table 6, the topic of expected composition is clearly explained. Furthermore, the test paper clearly specifies the composition's outline and marks instructions, including content, organization, grammar and appropriateness. In addition, the test paper also points out the scoring points of this section. Controlled composition seems to be beneficial to the improvement of scoring consistency.

Clear instructions can also contribute to the reliability of a test. In my questionnaire, the Question 2 is about whether the candidates think the instructions are clear. 90% of the candidates show that they think the instructions are clear. For example, in the Writing section mentioned above, students are informed of what the topical knowledge is about, how to deal with this section, and how the examiner will score this section. To take another example, in the Proofreading and Error Correction section, the test paper directly shows the candidates how to use the three ways of correcting in a right way (see Figure 10).

Figure 10: The Instruction for Proofreading and Error Correction (*2009 TEM-8 Test Paper* 2009)

Proofread the given passage on ANSWER SHEET TWO as instructed.

When ∧art museum wants a new exhibit, (1)___an_____

it never buys things in finished form and hangs (2)___forms____

them on the wall. When a natural history museum

wants an exhibition, it often build i. (3)____often____

This instruction is well against vagueness. With the instruction, students will not lose points because of using the wrong forms. On the other hand, being familiar with the format of the test helps test reliable as well. Question 1 in the questionnaire is about whether the candidates are familiar with the format of the test paper. 92% of the students indicate that they are familiar with the format since they have done a lot of practice and many model and previous tests before taking the actual test.

The scoring method is another matter that affects the test reliability. In TEM-8, the objective part takes up 40% of the total score while the subjective part takes up 60%. The objective part apparently has fixed answers to the items, and this part is done on an Answer Card, which is scored by a computer. Therefore, the score of the objective part is reliable. The subjective part is not as reliable as the objective part. However, the degree of reliability of this part is never low. There are four sections involved in the subjective part. The first section is Mini-Lecture in Listening Comprehension. The second section is Proofreading and Error Correction. These two parts both have fixed answers, thus they are highly reliable. The third section is Translation and the fourth section is Writing. For both sections, there are two certain grade score descriptions on the internet (Enfamily 2006). The scores are divided into five categories, each with specific descriptions of standard. As a whole, the score reliability is relatively high.

The four parts discussed above indicate that TEM-8 has consistency from occasion to occasion, controlled range of tested ability, clear instruction and familiar format, and relatively high score reliability. We can make the conclusion that TEM-8 has relatively high reliability.

3.2.2 Validity

To measure the extent of validity of TEM-8, three aspects of evidences should be analyzed – content validity, construct validity, and score validity.

As mentioned above, a specification of the skills or structures, etc. that a test is meant to cover is needed for judging whether it has content validity or not. In this essay, the specification of TEM-8 is the combination of the framework and the requirements of this test. This essay has analyzed in detail these two parts in **3.1.2 Implementation** section. In the framework of TEM-8, there are

specific details about the structure of the test, and in the requirements of TEM-8, detailed information about required skills is provided. TEM-8 is to test students' listening ability, reading ability, as well as writing ability. Different sections of the test can assess different abilities among the three. Listening Comprehension tests all the three abilities since the dictation is to test listening ability, the questions and Mini-Lecture passage are to test reading ability, and the gap-filling in Mini-Lecture is to test writing ability. Reading Comprehension and General Knowledge both test reading ability. Proofreading and Error Correction, Translation, and Writing all test both reading ability and writing ability. Therefore, TEM-8 is relatively content valid.

When discussing construct validity, the construct definitions should be analyzed first. There are two kinds of construct definitions, syllabus-based construct definitions and theory-based construct definitions. TEM-8 belongs to syllabus-based construct definitions. The *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000) is specifically set for English major students, and TEM-8 is set relying on this teaching syllabus. The syllabus is useful when we need to obtain detailed information on students' mastery of specific areas of language ability. There is a sample interpretation of construct validity in the Writing Section as shown in Table 4.

Table 4: The Interpretation of Construct Validity in Writing Section[9] (Zhou, 2003)

The Requirement in Teaching Syllabus	**Testing Ability**	**Testing Points**
Be able to use the basic translation theories to be preliminarily familiar with the comparison of English and Chinese languages, and acquire common translation techniques; Further with the comparison of English and Chinese languages, acquire the theories and techniques in English-Chinese and Chinese-English translations.	Translation Ability	1. Knowing the different ways and approaches to translation according to different topics, styles, occasions, situations and audience; 2. Expressing their ideas eloquently and expressively; 3. Conveying the information faithfully; 4. Understanding the differences between literal translation and meaningful translation; 5. Knowing the basic techniques in translation; and 6. Understanding the differences in usage, word order, grammatical structure and rhetorical devices between English and the mother tongue.

[9] The syllabus used in this table is the old syllabus. However, it can still well interpret the construct validity in TEM-8.

Since TEM-8 tests multiple abilities in the test tasks, it is unlikely that the measurement of any abilities of the three can be accurate. However, TEM-8 is a test assessing students' integrative ability, and the score reliability is relatively high, therefore, the inaccurate measurement of a single ability will not affect the validity of the whole test too much. Question 5 in the questionnaire is about whether the candidates think the result of TEM-8 can reflect their language ability or not. 12% of the students chose "absolutely", the majority, 64% of the students chose "maybe", 18% of them chose "no" and 6% of them chose "I don't know". From this data, we can find that the majority of the students indicate that the result TEM-8 can probably reflect their language ability. Of course, the test scores can never be considered absolutely valid since valid is just a degree of measurement and the test validation is an on-going process.

Question 6 in the questionnaire is about which sections the candidates think can measure their English language ability. 80% students chose Listening Comprehension, 64% students chose Reading Comprehension, 10% students chose General Knowledge, 24% students chose Proofreading and Error Correction, 90% students chose Translation, and 70% students chose Writing. It seems the content, construct and score of Listening Comprehension, Reading Comprehension, Translation, and Writing are considered reflecting a great degree of students' language ability, especially with the Translation section. In combination with the analysis earlier, we may have the conclusion that TEM-8 has great validity.

3.2.3 Authenticity

There are three categories in testing authenticity, input (material) authenticity, task authenticity, and layout authenticity. Within input authenticity, it further falls into three aspects, situation authenticity, content authenticity, and language authenticity. In this essay, the focus is on language authenticity and task authenticity.

All English major students receive education in English, including their learning material, learning syllabus, homework, classroom teaching, as well as tests. Although they are not native speakers, in the classroom, the teacher teaches in English, explains in English, and the students interactive in English. All the instructions are in English. It seems in their learning process, except for the translation part, Chinese language is never their language of instruction. Therefore,

the whole test paper in English, questions as well as introductions, seems to be very authentic when related to English majors' real learning life.

The tasks of TEM-8 have with great authenticity as well. One can relate all the sections in TEM-8 to real life. For **Listening Comprehension**, all the three parts of the task can be found in the real study process. In the academic world, classroom lectures by professors are very common for a non-native English student. Students always write down the main point or the key points of the lectures. This is the same in Mini-Lecture part. The Mini-Lecture part lets students hear the lecture once, and requires the students to take notes. After that, an answer sheet with the main structure of the lecture is handed out. The students are supposed to fill in some important information to complete the structure. The Interview and News Broadcast can also be found easily in real life. When we listen to a piece of interview or news, we hope we can find the main point of it. For **Reading Comprehension**, since everyone reads books, newspapers, articles and some other material, they are of the same function. The material provides information, and the readers look for the information they want from the material. For **Proofreading and Error Correction**, it is the most authentic since everyone can make mistakes in writing and they have to be corrected. For **General Knowledge**, since there are courses about English-speaking countries' general situation, English literature, American literature, and linguistic knowledge, this section is closely related to the real teaching syllabus. For **Translation**, all non-native English learners are familiar with the translation between English and their mother language. Furthermore, there are also translation courses according to the Teaching Syllabus. For the final section—**Writing**, it is also very close to the real learning process. Students are always required to write an essay after reading a book, listening to a lecture, and watching a movie. This is a way they can express their ideas and thoughts in paper, only in English.

In the questionnaire, Question 7 is about whether the candidates are familiar with the test items and techniques in the test, and 78% students chose "yes". The test designers design the test according to the Teaching Syllabus, and at the same time, all our courses and learning material are also designed according to the syllabus, as well as the classroom activities. To sum up, TEM-8 has relatively high authenticity.

3.2.4 Interactiveness

To measure the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task, the relationship between input and response should be analyzed. Brown and Hudson (1998: 653) divide the language assessment methods into three types—selected-response assessment, constructed-response assessment, and personal-response assessment. Davies (1990:38-39) provides a framework of discrete point—integrative, in which the characteristic of assessment is determined by four elements— input, nature of task, nature of response, and scoring. Bachman and Palmer (1996: 55) classify the test task into three types—reciprocal tasks, non-reciprocal tasks, and adaptive tasks. Combining the three ways of categorizing testing methods or tasks, here comes the framework of testing methods of TEM-8.

Table 5: The Framework of Testing Methods of TEM-8

The Title of Task	Nature of Input	Nature of Task	Nature of Response	Scoring	Interactiveness
Listening Comprehension A (Mini-Lecture)	Integrative	Integrative	Constructed	Discrete point	Reciprocal
Listening Comprehension B,C (Interview and News Broadcast)	Integrative	Discrete point	Selected	Discrete point	Non-reciprocal
Reading Comprehension	Integrative	Discrete point	Selected	Discrete point	Non-reciprocal
General Knowledge	Integrative	Discrete point	Selected	Discrete point	Non-reciprocal
Proofreading and Error Correction	Integrative	Integrative	Constructed	Discrete point	Reciprocal
Translation	Integrative	Integrative	Constructed	Integrative	Reciprocal
Writing	Integrative	Integrative	Constructed	Integrative	Reciprocal

All these sections are designed with integrative input. Except for Translation and Writing, all other sections are discrete point scoring. Objective parts are discrete point tasks with non-reciprocal interactiveness that requiring selected response, while subjective parts are integrative tasks with reciprocal interactiveness that requiring constructed response. In this table, we can find that most testing methods are integrative tasks, which can ensure relatively high interactiveness.

The characteristics of language test task consist of three aspects, language ability (including language knowledge, metacognitive strategies), topical knowledge, and affective schemata. Since the test takers have similar language ability and affective schemata, this section only focuses on the topical knowledge. Some test tasks that require certain topical knowledge may be easier for those who have that knowledge and more difficult for those who do not. However, for most English major students, they have equal knowledge of subjects outside of their own, such as politics, science, economy, culture and tradition. In the TEM-8 of 2009, the topics of four Reading Comprehension texts are travelling abroad, the baby boom and its impact, working wives, and hill climbing. They have covered politics, science, culture, tradition, and society. 52% of the students show that they are familiar with the topics in the test, while the rest of the students show they are not. This shows many students still lack large-scale topical knowledge. The *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000: 4) has pointed out that the students are supposed to be familiar with working in department of foreign affairs, education, economy, culture, science, and military. It also points out that the students should acquire large scale of knowledge and certain relative major knowledge. It is not only beneficial for better using English in working, but also important to develop integrative students. Therefore, there are certain relative courses set up according to the Teaching Syllabus. Therefore, the topical knowledge of TEM-8 is relatively interactive as well.

3.2.5 Practicality

TEM-8 is high in practicality. It is set up by the Ministry of Education, and the Higher Education Institution Foreign Language Teaching Supervisory Committee English Group is responsible for it both academically and organizationally. The group is composed of professors and experts from several top colleges and universities in the country.

With the use of computers to score the objective parts in an Answer Card, the examiners have removed certain burdens, and this method well avoids certain man-made mistakes, such as failing to see the wrong answers. It saves time, human resource and money.

3.3 Impact

Tem-8 is to assess and evaluate the actual performance of *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000) for senior students. Meanwhile, it assesses the teaching quality as well as the students' language ability. As a high-stakes test, it has certain impact on the test takers, the teachers, and the society and educational systems.

3.3.1 Impact on Test Takers

As mentioned earlier, there are three aspects of testing procedure that affect test takers—the experience of taking and preparing for the test, the feedback about their performance on the test, and the decisions made by the results of the test.

The Experience of Taking and Preparing for the Test

TEM-8 is a high-stakes nation-wide public examination, used as an achievement test as well as a proficient test to select and label different test takers within different levels. Preparing for the test costs students a lot of time, money, and energy.

According to the data of questionnaire (Question 9), 80% of the students took 0-3 months to prepare this test, and the remaining 20% of the students spent 4-6 months on it. Students have to spend several weeks or even months to prepare such high-standard examination. One of the students points out that, in fact, students have been preparing for the test since they begin learning the English language. TEM-8 is to assess students' language ability, however, the language ability is accumulated day after day, year after year. From the time of preparation for the test, the importance of TEM-8 is apparent. However, the preparation time and the time taking the test affect students' daily life. As mentioned earlier, 60% of the students state that the test taken in Semester 8 is not suitable. Although it is taken in the beginning of Semester 8, students already have a lot of things to do which are closely related to their future life in Semester 7. In

Question 15, 14% of the students think the preparation for TEM-8 influences the completion of their graduation paper (the first whole draft of the graduation paper is supposed to be handed in at the end of Semester 7). 30% of the students think the taking time of TEM-8 influence their hunting for jobs since many good jobs require the certification of TEM-8 when applying, however the late outcome stops students from getting the jobs as early as possible. 24% of the students, most of who are preparing for the Graduate Record Examination (taken in late January) in Semester 7, think the preparation for TEM-8 influence their English language learning. This puts the students under a certain degree of stress.

In addition to taking classes, students prepare the test themselves. They always buy material for preparation. As the Supervisory Committee has no official web site, they have no recommended material for students to prepare the test. What students can do is walk into the bookstore and buy some books or material they think may be helpful. There are six sections in the test, thus many publishers publish books or material for each section respectively. According to the data of the questionnaire (Question 10), 4% of the students did not buy any material for respective sections. They just bought model tests and previous tests as practice. 42% of the students bought material for one or two sections, 34% of them bought for three or four sections, and 20% of them bought for five or six sections for preparation. The material is not cheap at all. If a student buys material for all the six sections, adding model tests or previous tests, he will spend hundreds of RMB.

After students buy material for preparation, they have to spend time on the material. From the framework of TEM-8, we can see that in a controlled situation with limited time, students have to spend more than three hours on a complete test paper. It is not hard to imagine how much time the students would spend on practice. In addition, to increase the pass rate of TEM-8 and to comply with many students' request, some colleges and universities set up courses directly related with TEM-8. However, according to the questionnaire (Question 11), only 32% of the students think the courses are useful, and 56% of the students think the courses are not that useful. Some of the students who consider the courses useful state that the courses help them a lot, especially with Listening Comprehension and General Knowledge. Teachers can teach them better technique to do Listening Comprehension, and provide the most important material for General Knowledge so that students do not have to waste their time to find all relative material.

For those who think the courses are not that useful, some of them state that the courses just provide more practice rather than instructions, and some others state that the courses are not enough and the time is limited. Although not everyone is required to attend these courses, students who attend these courses spend more energy than those do not on these courses and the after-class homework.

The Feedback on Test Taker's Performance on the Test

The feedback on test taker's performance on the test is the score of the test in other words. In TEM-8, as mentioned earlier, the students get their scores and certifications from their school. There is a weakness that as mentioned earlier, there are no official answers of the test, therefore the students cannot get to know where they have made mistakes. It is not useful if students want to find out their strong points and weak points of language ability. Although only 64% of the students consider the test to reflect their language ability, experts design the test in this way for certain reasons. Since the test has high reliability, validity, authenticity, interactiveness and practicality, we may say that the test reflects students' language ability to a great degree. Therefore, the feedback on the test taker's performance on the test is relatively meaningful for the test takers.

The Decisions Made on the Basis of the Result of the Test

As a high-stakes test which is used as a proficient test, TEM-8 has great influence on students. Although the test in some ways has maleficial impact on students' daily life, it has some beneficial impact on students as well. In Question 14, about in what aspects TEM-8 helps students, 28% students chose "testing language ability", 54% chose "finding a job", 20% chose "learning English language", 30% chose "becoming more confident in English". This is especially true with "finding a job". As mentioned earlier, many good jobs need the certification of TEM-8 when being applied for. After having held the certification, the student adds one more strong point on his curriculum vitae. Since more than half of the test takers around the country cannot pass the examination, one having the certification has advantage when applying for a good job competing with one who does not.

However, consequently, the students feel stressed because of this test. TEM-8 is the highest standard for English major students as well as non- English major students. For some students,

the certification of TEM-8 is as important as the diploma. Every English major student attend in the examination and wish to pass it. In the Question 13 about the stress students feel, 22% students feel much stressed to take this test, 64% students feel some stress about the test, 12% students feel little stressed, while only 2% students feel no stress with taking the test. There are several sources of the stress. Firstly, as mentioned in **3.1.2 The Implementation of the Test** the framework of the test section, some stress is from the procedures of taking the test. Secondly, people always evaluate English majors' language ability with the certification. Maybe the interviewer of a company who does not really know English thinks that if the interviewee has the certification, he must have high level of English ability and he must be better than those do not have. Therefore, it has a symbolic meaning for English majors. Thirdly, there is some stress from the family. Parents wish their children can pass the highest standard examination of their majors and the parents will be proud of them if they pass the examination. The three sources are all realistic, therefore, when I ask the candidates whether they will take the next year's examination if they fail in the first year (Question 16), 76% of them chose "yes". The certification is so important for English major students.

3.3.2 Impact on Teachers

As a high-stakes test, teaching to the test is an unavoidable situation. As mentioned earlier, some colleges and universities set up courses directly related to TEM-8. Therefore, some teachers are asked to teach the courses. This essay has interviewed two teachers of a university in China to see teachers' general opinion of the test as well as the impact of the test on teaching structure.

The two teachers are respectively Teacher A and Teacher B. They are from the same university and teach the course directly related to TEM-8. The course is named "English Testing" which is set in Semester 7. The course directly tutors the six sections of the test, and since there are too many sections to prepare for just one teacher, two teachers are in charge of this course. For example, Teacher A is in charge of teaching Listening Comprehension, Reading Comprehension, and Writing, while Teacher B is in charge of teaching General Knowledge, Proofreading and Error Correction, and Translation. Teacher A has been teaching this course for three years while Teacher B has been teaching for four years.

The two teachers hold different points of view towards whether it is necessary to set up this course. Teacher A considers it necessary while Teacher B considers it unnecessary. However, both of them state that the practice in class helps students to be familiar with the format and instruction of the test, and students can be more skilful using techniques in different sections.

For the sources of material used in class, both teachers state that, they find the material together with another teacher (there are altogether four teachers teaching this course with two classes) from a variety of sources like foreign translations, textbooks, and model and previous test paper. They choose the material mainly for the reason that those kinds of material are typical and may reflect students' problem in each section. They are helpful for students to find the problems and thus to solve them.

Regarding to the timetable of the course, the two teachers also have different opinions. In Semester 7, the course "English Testing" is given to students once a week, with each class lasting for one and a half hour, and each teacher takes half semester. However, the difference of opinions is related to the different sections that the two teachers are in charge of. Teacher A is in charge of Listening Comprehension, Reading Comprehension, and Writing. The reading comprehension and writing practice is handed out as homework at the end of last class. Therefore, the teacher just needs to do listening comprehension in the class and comment on students' previous work on reading comprehension and writing. Teacher B is in charge of General Knowledge, Proofreading and Error Correction, and Translation. All the three parts are done in the class and commented immediately. Since there is much knowledge and many points of all the three sections should be referred to, an hour and a half of a class is enough for neither teacher nor students.

Both teachers state that the students should prepare for the test for about three to six months. However, for some students who are studying English very hard all the time, one month may be enough as what they need is to know what to be tested and how to arrange the time probably.

Referring to the validity of the test, both teachers show that they consider the test can really reflect students' language ability. TEM-8 is a comprehensive test, and generally, the score is a good reflection of the students' language efficiency.

The students consider TEM-8 as very important, and so do teachers, but only for students. Nowadays, it is one of the most important ways to tell the language efficiency of English majors. However, the test is of no importance to teachers. With good results, the teachers will not get reward, and with bad results, the teacher will not be punished either. In converse, the result may have something to do with the university. Good result like high pass rate and Distinction rate may help build up the fame of the university.

3.3.3 Impact on Society and Education Systems

Since TEM-8 is the highest standard of test evaluating the language ability of English majors, most people take it as a measurement to see whether an English major student is proficient enough. This kind of thought brings maleficial impact on students that they may feel very stressed, and the thought is not suitable for judging a person's proficiency just relying on a piece of paper as well. Some students may be good at taking test while some may be not. The test is not the only way to judge a person's language ability level. For example, in a job interview, in addition to the TEM-8 certification, the interviewer should consider other aspects as well. From the school academic report, the interviewer can find out the student's daily performance of study, whether he has made progress, whether he has worked hard enough, and what his strong points and weak points are. From the academic activity recording, the interviewer can see whether the student is good at language use. Furthermore, the interviewer can talk with the student directly, getting whatever information he needs to judge the student's ability. The society should judge a person's language proficiency with an integrative point of view.

The test has impact on the education system as well. On one hand, as mentioned earlier, teaching to the test is an unavoidable situation since TEM-8 is a high-stakes test. In addition to the courses directly related to the test (such as "English Testing"), many other ordinary courses are approaching the test. For example, in course of Listening, news broadcast and interviews are the common material, however, in recent years, the course gradually adds the part of note-taking. Although the note taking may be not the same as the form of Mini-Lecture in TEM-8, they share similarities. In another course, Senior Writing, the teacher asks the students to write compositions according to the requirement of Writing in TEM-8. Although teaching to the test is beneficial for

students who are going to take the examination, we hope in the meantime, the courses can really help promote the students' language ability.

Furthermore, testing is a good way to evaluate the quality of teaching. From the test, experts can analyze the actual performance of Teaching Syllabus, and measure students' language ability. In fact, the test and the Teaching Syllabus interact with each other. A good test can reflect the actual performance of the Teaching Syllabus, and in what aspects the Teaching Syllabus should be improved. Conversely, a better Teaching Syllabus can make better Testing Syllabus, thus to make better test to evaluate the Teaching Syllabus better. Based on the analysis of the test, experts can bring forward new requirement for students to push them onto a higher level of language ability, and new Teaching Syllabus for teachers to give better education to the students.

4. Conclusion

TEM-8, Test for English Majors Band 8, is a test to assess and evaluate the actual performance of *Higher Education Institution English Major English Teaching Syllabus* (Higher Education Institution Foreign Language Teaching Supervisory Committee English Group 2000) for senior students. It measures the quality of teaching and students' language ability as well.

Test usefulness is a good way to measure whether the test is suitable. Consistency from occasion to occasion, controlled range of tested ability, clear instruction and familiar format, and relatively high score reliability indicate that TEM-8 has relatively high reliability. High content validity, construct validity, and score validity indicate that the test has great validity. High language authenticity and tasks authenticity indicate that the test is relatively authentic. Integrative tasks and familiar topical knowledge ensure relative interactiveness in the test. Being easy to be put into practice, and saving time, human resource and money indicate that the test is high in practicality. In this standard, apart from the impact part, we may say that TEM-8 is relatively suitable for English majors.

The system of TEM-8 is clear and well organized. The test has the specific purpose, the implementation follows a strict procedure, the framework and requirements are very detailed, and the reform of the test is reasonable and meaningful. However, the system has many aspects of

impact on students, both maleficial and beneficial. With no official web site, students find it inconvenient to find some important information; the timetable of the test influences students' daily life, including hunting for jobs, preparing for the graduation paper, and applying for further education; the testing procedure makes students feel stressed; the high requirements need students to improve their language ability into a higher level; and the reform of the test provides more and higher requirements for students. The maleficial impacts need to be considered by the committee and ameliorated in further study.

The impact of the test on students is very great. From preparing for the test, to getting feedback of the performance of taking the test, and last receiving decisions by the result of the test, students get impact on their time, money, energy, reflection of language ability, future life.

The test has beneficial as well as maleficial impact on students. With the interaction of the Teaching Syllabus and the test, the improvement of the Teaching Syllabus is supposed to make a better test. Therefore, the test can bring more impacts that are beneficial for students and reduce the maleficial impacts at the same time. In this way, the test can be used better to evaluate the performance of the Teaching Syllabus as well as students' language ability.

References

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. & A. Palmer (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Bo, J. (2007). An Analysis of Authenticity in CET-4 and TEM-8. *Sino-US English Teaching*. Vol. 4, No. 2, 2007, 28-33.

Brown, H. (2004). *Language Assessment: Principles and Classroom Practices*. The United States of America: Pearson Education, Inc.

Brown, J. & Hudson, T (1998). The Alternatives in Language Assessment. *TESOL Quarterly*. Vol. 32, No. 4, 1998, 653-675.

Chang J., Zhang Y. & Wu Y. (2006). A Study of the First Implementation of 2004 Syllabus for TEM-8. Foreign Language and Their Teaching. Serial No. 208, 2006, No.7, 18-21.

Davies, A. (1990). *Principles of Language Testing.* New Jersey: Basil Blackwell Lte.

Enfamily (2006-06-29). *TEM-8 Grade Score Description.* Accessed May 10, 2010. <http://bbs.enfamily.cn/thread-10050-1-1.html >

Gennaro, K. (2006). Fairness and Test Use: The Case of the SAT and Writing Placement for ESL Students. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*. Vol. 6, NO. 2, 2006. Accessed May 8, 2010. <http://journals.tc-library.org/templates/about/editable/pdf/Di%20Gennaro%20Forum.pdf>

Higher Education Institution Foreign Language Teaching Supervisory Committee English Group (2000). 《高等学校英语专业英语教学大纲》[*Higher Education Institution English Major English Teaching Syllabus*]. Beijing: Foreign Language Teaching and Researching Press.

Higher Education Institution Foreign Language Teaching Supervisory Committee English Group (2005). 《专八考试新大纲》[*The Syllabus of Test for English Majors Band 8*]. Accessed May 10, 2010. <http://www.hjenglish.com/subject/tem/page/14872/?page=1>

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Lan Y. & Huang X. (2004). TEM-4 & TEM-8 Feedback on English Teaching and Problems Caused by Expanded Enrolment and the Corresponding Solutions. *Guangxi Teachers College Journal (Philosophy, Social and Science Edition)*. 2004, 112-116.

Li B., Wang A. & Wang X. (2007). Teaching Feedback Upon the Statistical Analysis of TEM-8 Results. *Journal of Taizhou University*. Vol. 29, No. 5, 2007,78-86.

Messick, S. (1993). Validity. [In] Linn, R. (ed.) *Educational Measurement.* Phoenix: The Oryx Press.

Miller, M. (1985). *Reliability and Validity*. Western International University. Accessed May 5, 2010. <http://michaeljmillerphd.com/res600_lecturenotes/Reliability_and_Validity.pdf>

OPTISM n.d.. *Test Reliability and Validity Defined*. Randy, L. Ohio. Accessed May 5, 2010. <http://cc.ysu.edu/~rlhoover/OPTISM/reliability_validity.html>

Popham, W. (2002). *Classroom Assessment: What Teachers need to know*. Boston: Allyn & Bacon.

Research Methods Knowledge Base (2006). *Reliability & Validity*. William M.K. Trochim. Accessed May 5, 2010. <http://www.socialresearchmethods.net/kb/relandval.php>

Shohamy, E. (1998). Critical Language Testing and Beyond. *Studies in Educational Evaluation*. Vol. 24, No. 4, 1998. 331-345.

Twenty-four English Website (2009-03-09). *2009 English TEM-8 Test Paper*. Accessed May 3, 2010. <http://www.24en.com/tem/dynamic/2009-03-09/106418.html>

Weigle, S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wells, C. & J. Wollack (2003) *An Instructor's Guide to Understanding Test Reliability*. University of Wisconsin. Accessed May 4, 2010. <http://testing.wisc.edu/Reliability.pdf>

Zhou, S. (2003). The Cohesion of Language Curriculum and Language Test—the Design and Implementation of TEM-8. *Foreign Language World*. No. 6, 2003 (General Serial No. 98), 71-78.

Appendices

Appendix 1: Questionnaire

Questionnaire

Hey everyone, I am working on a study about the impact of TEM-8 on English majors, and I know you have just gone through the test in March. So I will appreciate it if you can accomplish this questionnaire. Thank you for your time.

1. Are you familiar with the format of the test? If yes, how? If no, which section's format of the test is it that you are not familiar with?
   A. yes B. no

   _____

2. Do you think the instructions are clear? If not, in which sections?

   A. yes B. no

   _____

3. Do you think the test is difficult? And why?

   A. yes B. no

   _____

4. What do you think about the time limitation for each section?

   A. not enough B. just in time C. enough D. whatever

   Listening Comprehension _____

   Reading Comprehension _____

   General Knowledge _____

   Proofreading & Error Correction _____

   Translation _____

   Writing _____

5. Do you think the result of TEM-8 really reflects your language ability? Why?

   A. absolutely B. maybe C. no D. I don't know

   _____

6. Which part/parts do you think can measure your English ability? (you can choose more than one answer)
   A. Listening Comprehension B. Reading Comprehension

   C. General Knowledge D. Proofreading & Error Correction

   E. Translation F. Writing

G. None

7. Do you think the test items and techniques (gap filling, multiple choices, writing) are related to how you practice English in class and after class? If no, why?
   A. yes B. no

   _____

8. Are you familiar with the topics used in the test?

   A. yes B. no

9. How many months earlier did you start to prepare TEM-8 before the test?

   A. 0-3 months B. 3-6 months C. 6-12 months D. more than 12 months

10. There are six sections in the test, how many sections did you buy material for preparation?
    A. 0 B. 1- 2 C. 3-4 D. 5-6

11. Do you think the courses related to TEM-8 are useful for your preparation? If it is useful, in what aspects are useful? If not, why?
    A. very useful B. useful C. not that useful D. useless

    _____

12. Do you think the test taken in the eighth semester is suitable? And why?

    A. very suitable B. suitable C. not suitable D. I don't mind when

    _____

13. Do you find it stressful to take this test? And why?

    A. very much B. some C. little D. none

    _____

14. In what aspects do you think TEM-8 helps you? (you can choose more than one answer)
    A. testing language ability B. finding a job

    C. learning English language D. becoming more confident in English

    E. nothing

15. In what aspects do you think preparing for TEM-8 influences your daily life? (you can choose more than one answer, and you can write your own answers as well)
    A. completing graduation paper B. hunting for job

    C. learning English language D. nothing

    Other _____

16. If you fail in this test, do you want to take it next year? And why?

    _____

The Data of Questionnaire

		Percentage						
Item		**A**	**B**	**C**	**D**	**E**	**F**	**G**
1		92%	8%					
2		90%	10%					
3		88%	12%					
	L	12%	54%	34%	0%			
	R	74%	18%	6%	2%			
4	**G**	4%	24%	70%	2%			
	P	34%	34%	30%	2%			
	T	10%	42%	44%	4%			
	R	14%	48%	36%	2%			
5		12%	64%	18%	6%			
6		80%	62%	10%	24%	90%	70%	0%
7		78%	22%					
8		52%	48%					
9		80%	20%	0%	0%			
10		4%	42%	34%	20%			
11		6%	32%	56%	6%			
12		2%	28%	60%	10%			
13		22%	64%	12%	2%			
14		28%	54%	20%	30%	16%		
15		14%	30%	24%	24%			
16		76%(yes)	12%(no)	12%(maybe)				

Appendix 2: Interview

Questions:

1. How many years have you been teaching this course?
2. What sections of the test are you teaching?
3. Where do you get material for each section for students? And why do you choose them?
4. Do you think the timetable of this course is suitable? Why?
5. Do you think it is necessary to set up this course? Why?
6. For how long do you think the students should prepare for the test?
7. Do you think TEM-8 can really reflect students' language ability?
8. Do you think TEM-8 is important? In what aspects? (for students/ teachers/ faculty)

Appendix 3: Specifications for the TEM-8 (Excerpts)

The requirement of the test

Listening Comprehension

(a) Students should be able to follow English dialogues and speeches in company. (b) Students should be able to follow special coverage about politics, economy, culture and education, and technology within foreign media, such as VOA, BBC, and CNN. (c) Students should be able to follow general lectures about politics, economy, history, culture and education, language and literature, and science, as well as the questions and answers after the lectures.

Reading Comprehension

(a) Students should be able to understand articles about social discussion, politics and book analysis in general English or American newspapers or magazines. They should know the gist and general idea of the article, as well as tell the facts and details within the article. (b) Students should be able to read common biographies and literary works, understanding the literary meaning as well as the implicit meaning. (c) Students should be able to analyze the concept, structure, language skill and rhetorical devices within those types of articles mentioned above. (d) Students should be able to adjust their own reading speed during the test.

General Knowledge

(a) Students should have a basic idea about main English-speaking countries' geography, history, current situation, and cultural traditions. (b) Students should have acquired basic English literature knowledge. (c) Students should have acquired basic English linguistic knowledge.

Proofreading and Error Correction

The students should be able to identify speech disorders in a short paragraph with the language knowledge of grammar, rhetoric and sentence structure. In addition, they should be able to provide the correct alternative.

Translation

(a) The Chinese-to-English part requires students to be able to use the theory and techniques of Chinese-to-English to translate an argumentative essay, a narrative essay, and introductions about national conditions in China's newspapers and magazines, as well as excerpts of common literary works. The reading speed for this part is about 250 to 300 characters per minute. In addition, the translation should be strictly true to the original version, and the language should be fluent.

(b) The English-to-Chinese part requires students to be able to use the theory and techniques of English-to-Chinese to translate articles about politics, economy, history, culture and other aspects in English or American newspaper or magazines, as well as excerpts of original texts of literature. The reading speed for this part is about 250 to 300 words per minute. In addition, the translation should be strictly true to the original version, and the language should be fluent.

Writing

The students should follow the title and requirements given by the test paper and write an expository essay or argumentative essay with about 400 words. The essay should be written with fluent language, suitable words, reasonable style, and strong persuasion.