Assessing the Test Usefulness

A Comparison Between the Old and the New College English Test Band 4 (CET-4) in China

Lan Chen
Kristianstad University College
The School of Teacher Education
English IV, Spring 2009
D-essay in English Didactics
Tutor: Carita Lundmark

TABLE OF CONTENTS

1 Introduction

The College English Test (CET), one of the most pervasive English tests in China, has received much attention both from institutions of higher education and from educational departments concerned, greatly facilitating English teaching and learning since its introduction in 1980s. Widely accepted by the society, the CET-4 (Band 4 or Level 4) and CET-6 (Band 6 or Level 6) have served as one of the preconditions for the personnel departments at various levels to take on college graduates. In this way they have produced certain social benefits. At the same time, due to its large scale and extensive influence on college students both academically and psychologically, the test has been heatedly discussed in terms of its test contest and thus been under constant changes since then. Starting from 2005, the CET tests have been reformed, first in the scoring system, and later in the contents. Compared with the old test, the new system, concerned with the communicative skills of students, claims to better reflect the English proficiency of the college students, and therefore can greatly promote the implementation of college English teaching program as well as improving the teaching of college English as well.

This essay intends to take a closer look at the new system, and provide a basis for further study of the CET-4 trended towards a more communication-oriented test.

1.1 Aim

This paper is concerned with the newly reformed national English test for Chinese college students, called the College English Test (CET) Band 4 (or Level 4). By comparing the test before and after it was reformed, there will be a close examination with regard to the aspects of test reliability, construct validity, authenticity, interactiveness, impact and practicality. With an extra focus on how vocabulary and grammar are tested, the paper aims to investigate the extent to which the new system is considered useful and how effective it is in testing vocabulary and grammar.

1.2 Scope

This essay mainly looks into the six qualities of test usefulness according to the framework proposed by Bachman and Palmer (1996). The discussion mainly involves the contents of the tests and the way in which scores are reported. Aspects such as test-takers, the scoring of items and interpretation of scores will not be included in the present essay. More specific information in terms of the scope will be given at the beginning of each section of the analysis and discussion part.

1.3 Material

The official website, called the College English Test Band 4 and Band 6, provides the majority of the materials regarding the new CET-4 test that are to be analyzed and discussed. These materials include the new specifications, and the sample test of the new system. As the materials of the old system cannot be accessed from the official website, the Google website provides entries to those materials, including the old specifications and the sample test.

The sample test of the new system is the one released by the National CET-4 and CET-6 Commission together with the specifications. The old sample test for discussion is selected randomly from past test papers, and in this case, it is the January 2002 test paper. The individual items / questions will also be singled out from these sample tests. A detailed description of both tests can be found in section 3.1.

The results from previous surveys on the feedback from students, teachers and employers about the test will be analyzed and discussed. The reason why previous surveys are used is that in order to get reliable results, at least three parties will be involved, i.e. students, teachers and employers. The processes of distributing and collecting questionnaires as well as conducting interviews would take a long time, and is not possible within the time-frame afforded by this essay. It should be noted that as the three studies were conducted after the reform of the CET-4 test, the relevance of the survey might be questioned since they were done years ago. Undoubtedly, there could be changes in the results if they were done at

present. However, given the fact that the changes brought about by such a large-scale test require time, there might not have been dramatic modification of the teaching and consequently little change in students' performance on the test as one might assume. After all, the effect of such a large-scale test takes time to show. Hence, their results, in the main, are considered to reflect the general situation before the reform.

1.4 Method

To begin with, there will be an elaborative comparison between the old and new tests concerning their contents, together with the sample tests in order to investigate the extent to which the test is useful, in terms of test reliability, construct validity, authenticity, interactiveness, impact and practicality, respectively. At this stage, previous studies will be elicited. The data and results from the studies will be closely examined and discussed in order to find out to what extent the reformed test has impact on the society and the people involved.

Secondly, there will be a close examination into the reformed sample test, aiming to find out how abilities of grammar and vocabulary are tested and how effective the testing is. Items are selected from the sample tests for further analysis and discussion at this stage.

2 Theoretical background

With 15 years of development, there has been prolonged, extensive and profound research on the CET tests in China. A prominent example is the 3-year study on the validity of the test, starting from October 1992, conducted by the National CET-4 and CET-6 Commission in China and the Centre for Applied Language Studies (CALS) of University of Reading in Britain. The research on the CET tests is believed to have fostered innovation in classroom teaching and learning, generated the shift of focus from grammar to communication, and contributed to the enhanced comprehensive language ability of college students in China.

Apart from the study on the validity of the CET tests (Yang & Weir 1998; Miao 2006), research has been conducted that demonstrates the washback effect of the CET-4 test (Shao

2006), and its authenticity compared with the TEM-8 (Test for English Majors Band 8) (Bo 2007). Other studies have pointed at existing problems of the CET tests (Guo 2006a) and still others have looked into their future (Guo 2006b).

In this section, the six components of test usefulness will be defined and elaborated first, based on the framework proposed by Bachman and Palmer (1996), followed by theories on grammar and vocabulary testing, before, although not in this section, a profound comparison of the old and new testing systems is conducted.

2.1 The framework of test usefulness

Much previous research of various tests has based their discussion on Bachman and Palmer's framework of test usefulness (1996:18) (see Figure 1), which is considered as an important element in designing and developing a language test. According to Bachman and Palmer (1996:18), a model of test usefulness should include such qualities as reliability, construct validity, authenticity, interactiveness, impact and practicality.

Usefulness = Reliability + Construct validity +
Authenticity + Interactiveness + Impact + Practicality

Figure 1: A graphic representation of test usefulness from Bachman & Palmer (1996:18)

2.2 Test reliability

Test reliability refers to the consistency of scores on a test despite the varied occasions in which the test is administered. Bachman and Palmer (1996:19-20) highlight that reliability can be considered as a function of the consistency of scores from one set of tests and test tasks to another (see Figure 2).

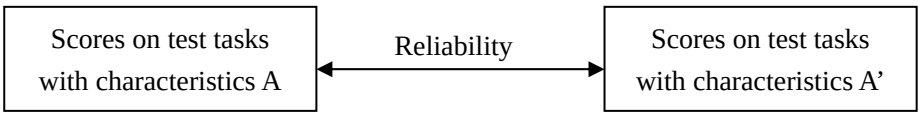

Figure 2: A graphic representation of test reliability from Bachman and Palmer (1996: 20). The double-headed arrow is used to indicate a correspondence between two sets of task characteristics (A and A') which differ only in incidental ways.

Due to the differences in the exact content being assessed on the alternate forms, environmental variables, such as fatigue, student error in responding, or even the lighting in the exam room, no two tests will consistently produce identical results (Wells & Wollack 2003). This is true regardless of how similar the two tests are. In fact, even the same test administered to the same groups of students will result in different scores. This being the case though, it does not imply that we can never have complete trust in any set of test scores. Hughes (2003:36) states the following:

> What we have to do is construct, administer and score tests in such a way that the scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time. The more similar the scores would have been, the more reliable the test is said to be.

That is to say, the highly reliable score ought to be "accurate, reproducible and generalizable to other testing occasions and other similar test instruments" (Ebel & Frisbie 1991: 76).

An important reason to be concerned with reliability is that it is a forerunner to test validity. That is, if test scores cannot be assigned consistently, it is impossible to conclude that the scores accurately measure the domain of interest. Ultimately, validity is the aspect about which we are most concerned. However, formally assessing the validity of a specific use of a test can be a laborious and time-consuming process (Wells & Wollack 2003). Therefore, reliability analysis is often viewed as a first-step in the test validation process. If the test is unreliable, one need not spend the time investigating whether it is valid–it will not be. If the test has adequate reliability, however, then a validation study would be worthwhile.

2.2.1 How to make tests more reliable

An approach to quantify the reliability of a test is the reliability coefficient, which involves complex formulae and for practical reason, this will not be a concern of this essay. However, researchers do suggest that a test can be made more reliable via technical approaches as follows (Hughes 2003:44-50):

1. Enough samples of behavior should be taken. The length of the test should be such that it

contains enough items which can well represent test-takers' language ability while avoiding the situation where candidates become so bored or tired that the behavior they exhibit becomes unrepresentative.

2. Candidates should not be allowed too much freedom on choosing test items; otherwise it is likely that there is great difference between the performance actually elicited and the performance that would have been elicited had the test been taken on another occasion.
3. Test items should be unambiguous. In other words, the meaning of test items should be presented clearly so that there will not be misunderstanding by the candidates or an unanticipated answer.
4. Clear and explicit instructions should be provided.
5. Tests should be well laid out and perfectly legible.
6. Effort should be made to ensure that candidates are familiar with the format and testing techniques, by distributing sample tests in advance, to prevent them from spending much time trying to understand what they are supposed to do.
7. Effort should be made to ensure scorer reliability by means of adopting items that permit scoring to be as objective as possible and that make comparisons between candidates as direct as possible (and this reinforces the suggestion that candidates should not be allowed too much freedom). There are also other means such as providing a detailed scoring key, training scorers, prior agreement of acceptable responses and appropriate scores, identifying candidates by number instead of name, and employing multiple, independent scoring especially where testing is subjective.

2.3 Test validity

Test validity pertains to the degree to which the test actually measures what it claims to measure. It is also the extent to which interpretations made on the basis of test scores are appropriate and meaningful. According to Hughes (2003:26), a test is considered to be valid if it measures accurately what it is intended to measure. If test scores are affected by other abilities rather than the one we want to measure, they will not be the satisfactory interpretation of the particular ability.

Language tests are created in order to measure a specific ability, such as ‘reading ability’, or ‘fluency in speaking’, which is referred to as a **construct**, on which a given test or test task is based which is used for interpreting scores. The term **construct validity** is therefore used to refer to the general notion of validity, and the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s) that we want to measure.

Bachman and Palmer argue that when test scores from language tests are interpreted as indicators of test takers’ language ability, “we need to demonstrate, or justify, the validity of the interpretations made of test scores.” (1996:21)

Content validity is one type of evidence which demonstrates that a particular interpretation of test scores is justified. A test is said to have content validity if its content constitutes a representative sample of the language skills, structures and so on, with which it is meant to be concerned. Moreover, the sample is expected to be representative so that it appeals to the purpose of the test. Therefore, a specification of the skills or structures, etc., that the test is meant to cover is needed for the purpose. The specification will provide the test constructor with the basis for making a principled selection of elements for inclusion in the test (Hughes 2003:27). A comparison of test specification and test content is the basis for judgments as to content validity.

The second form of evidence of a test’s construct 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, referred to as criterion-related validity, which is further divided into concurrent validity and predictive validity.

Apart from the test items, the way in which the responses are scored should also have validity. Scores are the basis on which inferences about a construct definition, or specific language ability, are made. Also, it is these scores that test users will make use of. Bachman and Palmer state that “[b]ecause test scores are commonly used to assist in making decisions about individuals, the methods used to arrive at these scores are a crucial part of the measurement process […], [which] play a key role in insuring that the test scores are reliable and that the

uses made of them are valid… ” (1996:193).

Bachman and Palmer point out that the type of score to be reported is determined by the construct definition. There are three types of reporting scores, namely, a single composite score, a profile of scores for different areas of language ability, and a combination of both (1996:194).

A composite score is a single score that is the sum or average of the scores from different parts of a test, or from different analytic rating scales. The test developer can use the raw scores or ratings, or, if some components are identified as more important than others, weight the importance of the components and multiply them by a number greater than one. A composite score can either be a compensatory one or a non-compensatory one. A compensatory composite score can be derived when an individual is assumed to have high levels in some of the areas of language ability to be tested and low levels in other areas. In this situation, a sum or average of component scores might balance out high scores and low scores. A non-compensatory composite score is adopted when there are high and low scores achieved in several areas of language ability, only the lowest score is used, which demonstrates the minimum level of mastery in several areas of language ability. In this case, the high score does not compensate for the low score.

The second way of reporting scores is one where a profile of scores corresponding to different areas of language ability is reported. The third way is a combination of a single composite score and a profile of scores that present the performance in each area of language ability to be tested.

2.3.1 How to make tests more valid

Hughes (2003:33-34) recommends such ways as the following to make a test valid:

> 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 responses 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.

In the development of tests, especially a high-stakes test, where significant decision about the individual is to be elicited from the results, there is an obligation for test developers to carry out a valid exercise before the test is in operation. However, it is worth noting that test validation is an on-going process and that the interpretations we make of test scores can never be considered absolutely valid (Bachman & Palmer 1996:22). Therefore, full validation is unlikely to be possible.

2.3.2 The relationship of reliability and validity

The primary purpose of a language test is to provide a measure that can be used as an indicator of an individual's actual ability in this language. The two qualities are thus essential to the usefulness of any language test (Bachman & Palmer 1996:23). Cumming and Mellow (1995:77) point out that "validity cannot be established unless reliability is also established for specific contexts of language performance". That is to say, test validity is a requisite to test reliability. If a test is not valid, then reliability is moot. In other words, if a test is not valid there is no point in discussing reliability because test validity is required before reliability can be considered in any meaningful way. Likewise, if a test is not reliable it is also not valid (*Test Reliability and Validity Defined* n.d.).

2.4 Test authenticity and interactiveness

Two elements that are crucial but often neglected by research in the test usefulness framework are authenticity and interactiveness (see Figure 1).

2.4.1 Authenticity

A key element in the test usefulness framework is the concept of **target language use (TLU) domain**, which is defined 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". A TLU task is an activity that an individual is engaged in by using the target language, so as to achieve a particular goal or objective in a particular situation (Bachman & Palmer 1996:44).

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) (see Figure 3). Though having not been discussed in many books, it is considered as a critical quality because it relates the test quality to the domain of the TLU task and provides a measure of the correspondence between the test task and the TLU task. Authenticity "provides a means for investigating the extent to which score interpretations generalize beyond performance on the test to language use" (Bachman & Palmer 1996:23-24).

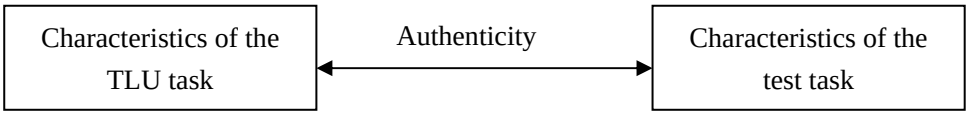

Figure 3: Authenticity (Bachman and Palmer 1996:23)

For example, in tests which examine communicative ability, the test construct must facilitate communication tasks which closely resemble the situations a test-taker would face in the TLU domain, so that they are more authentic. In fact, most language test developers implicitly consider authenticity in designing language tests (Bachman and Palmer 1996:24).

In attempting to design an authentic test task, the critical features that define tasks in the TLU domain are firstly identified. This recognition serves as a framework for the task characteristics. Test tasks that have these critical features are then designed and selected.

In a language test, authenticity is sometimes distantly related with real communicative tasks by carrying out a series of linguistic skills rather than genuine operational ones for reliability

and economy (Carroll 1980:37). A language test is said to be authentic when it mirrors as exactly as possible the real life non-test language tasks. Testing authenticity falls into three categories, which are input (material) authenticity, task authenticity and layout authenticity. Input authenticity can further be subdivided into situational authenticity, content authenticity and language authenticity.

2.4.2 Interactiveness

Interactiveness is another important element in the test usefulness framework proposed by Bachman and Palmers, which refers to "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (Bachman & Palmer 1996:25). Specifically, individual characteristics, i.e. the test-taker's language ability (including language knowledge and strategic competence, or metacognitive strategies), topical knowledge and affective schemata, which are engaged in a test, may influence the candidate's performance on the test (see Figure 4).

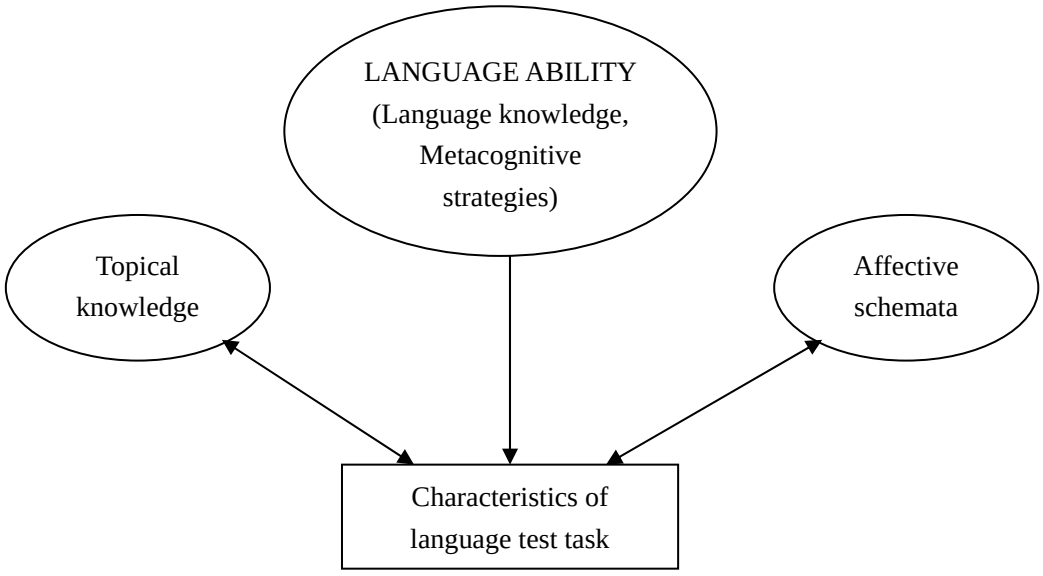

Figure 4: Interactiveness (Bachman & Palmer, 1996:26)

The double-headed arrows in Figure 4 represent the relationship, or interaction between an individual's language ability, topical knowledge, affective schemata and the characteristics of a test task. Due to these individual differences, the question is always how we could give each

test-taker a fair chance. Bachman and Palmer (1996:29) further highlight that for a test task to show a high level of interactiveness depends on its degree of correspondence with construct validity. Thus the importance of well-defined test-taker characteristics and the construct is clear and self-evident (see Figure 5). Otherwise, it is difficult to infer language ability based on an examinee's test performance when the test task does not demand that their language knowledge is used, despite a high level of interaction (Bachman & Palmer 1996:24).

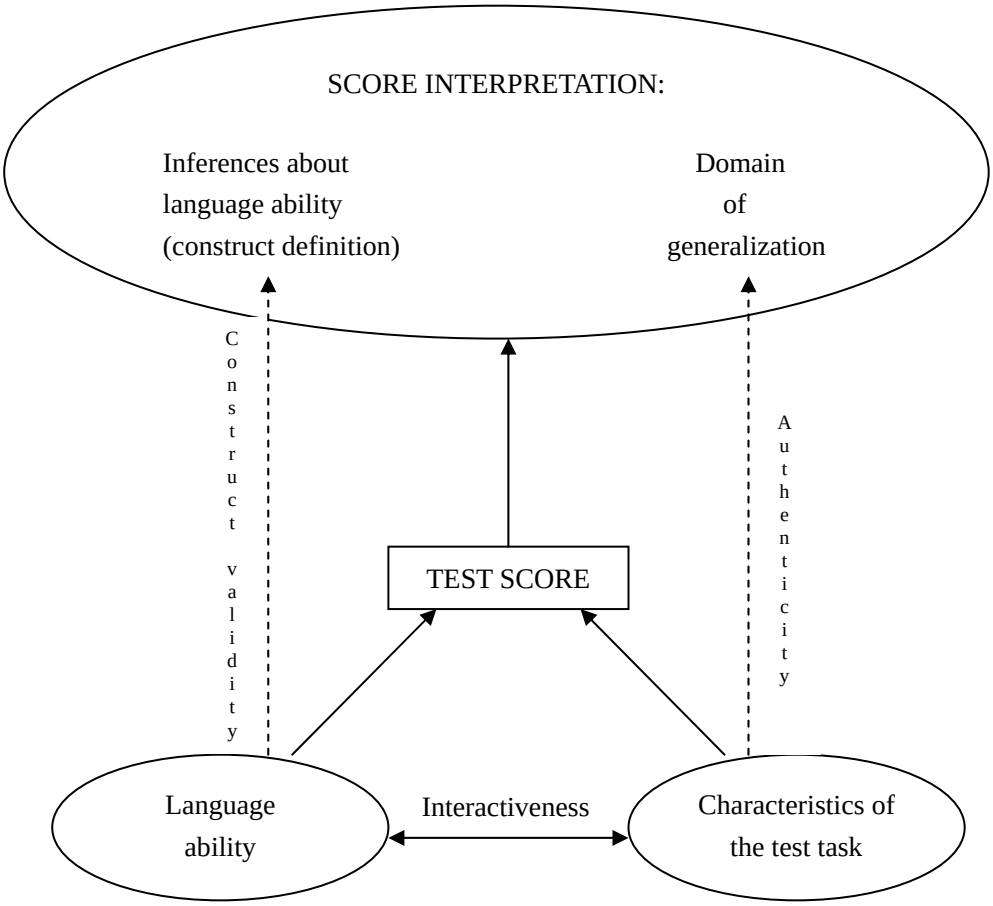

Figure 5: Authenticity and interactiveness and their relationship with construct validity

2.4.3 The distinction between authenticity and interactiveness and their relationship with construct validity

As is shown in Figure 5, both authenticity and interactiveness are linked inextricably to construct, so validity is first required to clearly establish the distinction between the two notions. Authenticity pertains to the correspondence between the characteristics of a test task and those of the TLU task, and is thus related to the traditional notion of content validity. It is thus highly dependent on the extent to which test materials and conditions replicate the TLU

situation (McNamara 2000:43). In the case of interactiveness, it indicates the interaction between the individual and the test task (of the test or TLU). That is, it is the degree of the test-taker's involvement when they are solving questions which assess their language competence, background knowledge, and affective schema.

2.5 Impact and practicality

Impact can be defined broadly in terms of the various ways in which test use affects society, an educational system at a macro level, and the individuals within these from a micro level (Bachman & Palmer 1996:39). Impact can be presented in Figure 6 below.

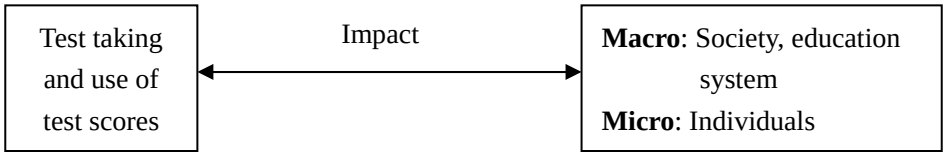

Figure 6: Impact (Bachman & Palmer 1996:30)

2.5.1 Washback

When we deal with the notion of impact, we must first get to know an important aspect of impact referred to as the "washback" (Bachman & Palmer 1996:30) or "backwash" (Hughes 1989:1). The concept pertains to the effect of testing on teaching and learning, and can be beneficial or harmful. An example of harmful washback is if a test includes no direct spoken component, it is possible that the skill of speaking will be downplayed or ignored completely in the classroom, to the ultimate detriment of the candidate's ability in that area, while the course objective is meant to train them in the comprehensive language skills (including speaking). "Teaching to the test" is an inevitable reality in many classrooms, and not only on those courses which aim to specifically prepare candidates for a particular exam. It is, therefore, important to ensure that the test is a good test, in order that the washback effect is a positive one.

2.5.2 Impact on test takers

Test takers can be affected in terms of three aspects (Bachman & Palmer 1996:31). First, the

experiences of preparing for and taking the test have the potential for affecting those characteristics of the test takers. For example, when a high-stakes nation-wide public test, such as the one being discussed in this paper, is used for decision making, teaching may be focused on the specifications of the test for up to several years before the actual test, and the techniques needed in the test will be practiced in class. The experience of taking the test itself can also have an impact on test-takers, such as their perception of the TLU domain. Secondly, the types of feedback which test-takers receive about their test performance are likely to affect them directly. Hence, there is a need to consider how to make feedback as relevant, complete and meaningful as possible. Finally, the decisions that may be made about the test takers on the basis of their test scores may directly affect them. In order for a fair test use to happen, test developers need to consider the various kinds of information, including scores from the test, which could be used in making the decisions, as well as their relative importance and the criteria that will be used.

2.5.3 Impact on teachers

In an instructional program the test users most directly affected by test use are teachers. In many occasions teaching to the test is found unavoidable. However, if a test is low in authenticity in the way that teachers feel what they teach is not relevant to the test, the test then could have harmful washback on instruction. To prevent this kind of negative impact on instruction, it, again, should be ensured that the test is a good one in order that the washback is a positive one.

2.5.4 Impact on society and educational system

Bachman (1990:279) points out that “tests […] are virtually always intended to serve the needs of an educational system or of society at large”. The very acts of administering and taking a test imply certain values and goals, and have consequences for society, the educational system, and the individuals in the system. This is of particular concern with high-stakes tests, which are used to make major decisions about large numbers of individuals (Bachman & Palmer 1996:34).

Shohamy (1998) further emphasizes the impact of tests on the society by putting forward the idea of **critical language testing**. She argues the following:

> "[…] 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."

This implies that language tests are not merely intended to fulfill curricular or proficiency goals as is previously defined, but have wider social and political implications as well.

2.5.5 Practicality

Practicality is defined 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" (Bachman & Palmer 1996:36) (see Figure 7). The resources required are specified as three types: human resources, material resources and time (Bachman and Palmer 1996:36-37). A practical test is one whose design, development, and use do not require more resources than are available.

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

If practicality$\geq$1, the test development and use is practical.
If practicality$\leq$1, the test development and use is not practical.

Figure 7: Practicality (from Bachman & Palmer 1996:36)

Of the six qualities in Bachman and Palmer's framework of test usefulness, practicality holds a great deal of importance in high-stakes testing contexts (such as a large-scale placement test[1]) (Gennaro 2006). Of course, all six qualities are relevant for test fairness, but practicality is a particular concern if it is given a disproportionate amount of weight compared to the other five components. High-stakes tests require a great deal of resources and, for this reason, are often considered costly and time-consuming. It is not surprising, therefore, that some test users may search for ways to avoid less practical performance tests if they believe other tests

[1] A placement test pertains to the test that are intended to provide information that will help to place students at the stage or in the part of the teaching programme most appropriate to their abilities (Hughes 2003:16).

can serve the same purpose. (Gennaro 2006). In short, the specific resources required will vary from one situation to another, as will the resources that are available (Bachman & Palmer 1996:40).

2.6 Testing grammar and vocabulary

Traditionally, the test of grammar and vocabulary has been considered by language teachers and testers as an indispensable part in any language tests since control of grammatical structures and a certain amount of word store are seen as the very core of language ability. Some large-scale proficiency tests, for example, retain a grammar and vocabulary section partly because large numbers of items can be easily administered and scored within a short period of time. In addition, as there are so many grammatical and lexical elements to be tested, it is impossible to cover them in any one version of the test, such as writing. It is therefore an advantage of the grammar and vocabulary test as there can be many items (Hughes 2003:172-179).

However, there has been a shift towards the view that since it is language skills that are usually of interest, then it is these which should be tested directly, not the abilities that seem to underlie them (Hughes 2003:172). There are two reasons for this change. For one thing, one cannot accurately predict mastery of the skill by measuring control of what is believed to be the abilities that underlie it. For another, the washback effect of tests which measure mastery of skills directly may be thought preferable to that of tests which might encourage the learning of grammatical structures in isolation, with no apparent need to use them. As Rea-Dickins (1997:93) mentions, it is unnecessary to test grammar as distinct forms but better to reflect it in some skill-based tests such as reading and writing; and it could be also conducted in another way, which grammar should be tested in an integrative way rather than simply be put into limited items in decontextualised single sentences. As a result, absence of grammar and vocabulary component has been seen in some well-known proficiency tests (Hughes 2003:172).

Vocabulary, which is embedded, comprehensive and context dependent in nature, plays an

explicit role in the assessment of learners' performance (Read & Chapelle 2001). The best way to test people's vocabulary is to use various ways to test either the basic meaning of a word, or its derived form, its collocations or its meaning relationship in a context. Nation (1990, as cited by Schmitt 2000:5) gives a systematic list of competencies which has come to know as types of word knowledge, which are 1) the meaning(s) of the word, 2) the written form of the word, 3) the spoken form of the word, 4) the grammatical behavior of the word, 5) the collocations of the word, 6) the register of the word, 7) the associations of the word and 8) the frequency of the word. These word knowledge types decide the meaning of vocabulary acquisition. Hence, if we want to analyze the construct validity of vocabulary items in the new CET-4 test, the key element is whether the meaning sense is typical way of usage in the academic context learners are in at present and a specific career-related context in the future. Schrutt (1999:192) points out that "[a]lthough any individual vocabulary item is likely to have internal content validity, there are broader issues involving the representativeness of the target words chosen".

3 Analysis and discussion

The usefulness of the old and new CET-4 tests will be analyzed and compared here. The items used in the discussion are elicited from the two sample tests, i.e. the sample test released by the National CET-4 and CET-6 Commission with the new specifications, and the January 2002 test paper. Focus is placed on the content of the tests, with reference to the specifications.

Before the comparison begins, the context of the test will be introduced. Then an effort is made to illustrate the old and new framework since the framework is where the major difference between the two tests resides. In the last subsection, there is an elaborative discussion on the testing of grammar and vocabulary in the new test.

3.1 The CET-4 context

Due to the huge discrepancy between Chinese and English and the worldwide popularity of

English, the English language has become an issue that has aroused increasing attention from the public and educational institutions. A tremendous amount of money and resources have been spent on research concerning improving Chinese students' English proficiency. In the job market, knowledge of English is considered an inevitable requirement in order to get a satisfactory job.

The old CET test has its place in the old social conditions and educational system in that the traditional method focused on reading and writing abilities, and the communicative function was not a prominent demand. However, with the reform in the educational system, the communicative aspect of the language has been called for. The CET test, therefore, has gone through several major changes, first in 2005 when a new scoring system was adopted, and then in 2006 when the test contents were significantly adjusted, whereby the direct test of vocabulary and grammar was removed and the listening and oral abilities are greatly emphasized.

In setting the content, the CET-4 uses the success of TOEFL for reference which is mainly about events and activities that happen on a university campus, covering aspects ranging from college life to students' knowledge structure, such as western customs and culture, science, technology and so on. The listening part, in particular, is characterized by this (Long Conversations, for example). Even the scoring system of the CET test bears resemblance to that of the TOEFL.

3.1.1 Test frameworks

The new system contains four parts, i.e. Listening, Reading, Cloze/Error Correction and Writing. Except for the Cloze/Error Correction part, the other three parts include subsections. Various testing techniques are adopted, such as multiple choice questions, Banked Cloze[2], Short Answer Questions. As far as score is concerned, the listening and reading parts both take up to 35% of the total score, respectively, while writing and cloze/error correction account for 20% and 10% respectively (see Table 1).

[2] Banked Cloze includes a short passage where there are 10 blanks that require candidates to fill in with the appropriate words from 15 choices (see Figure 9 on page 27).

Table 1: The framework of the new CET-4 test

Part	Test item			Format	Number of items	Percen -tage	Time (min.)
I	Listening Compre-hension	Dia-logue	Short dialogues	Multiple choice questions	8	15%	35
			Long dialogues	Multiple choice questions	7		
		Passages		Multiple choice questions	10	20%	
				Compound dictation	11		
II	Reading	Reading in Depth		Multiple choice questions	10	25%	40
				Banked Cloze / Short answer questions	10		
		Skimming and Scanning		True / false statements + gap filling	10	10%	
III	Cloze	Cloze / Error correction		Multiple choice questions	20	10%	15
IV	Writing & Transla-tion	Writing		Passage writing	1	15%	30
		Translation		Translation from Chinese to English	5	5%	5
Total	4				92	100%	125

The old system comprised five parts, i.e. Listening, Reading, Vocabulary and Structure, Cloze/Translation, and Writing. Among the five parts, only the listening part contained subsections. Most parts adopted the multiple choice format (70%-90%). As far as score is concerned, reading formed the dominant proportion of 40% in the total score, followed by the listening part, which constituted 20%. The part of Vocabulary and Structure shared the same percentage with Writing. Finally, Cloze/Translation made up the smallest percentage of 10% (see Table 2).

By comparing the layouts of the two tests, it is revealed that 60% of the test items in the reformed system is inherited from the old one, and the rest 40% is newly added. Also, varied types of questions are augmented, resulting in an increased amount of test items and the demand for an enhanced speed in doing the test.

Table 2: The framework of the old CET-4 test

Part	Test item		Format	Number of items	Percen - tage	Time (min.)
I	Listening Compre-hension	Dialogues	Multiple choice questions	10	10%	20
		Passages / Spot dictation / Compound dictation	Multiple choice questions / gap-filling	10	10%	
II	Reading Comprehension		Multiple choice questions	20	40%	35
III	Vocabulary and Structure		Multiple choice questions	30	15%	20
IV	Cloze / Translation from English to Chinese		Multiple choice questions / translation	20	10%	15
V	Writing		Passage writing	1	15%	30
Total	5			91	100%	120

The reformed test has excluded two question types, i.e. Vocabulary and Structure, and Translation from English to Chinese, while Cloze and Writing remain exactly the same as the old test.

The percentage of Listening Comprehension is increased from 20% to 35%. The three techniques, i.e. Short dialogues, Passages and Compound Dictation, are taken from the old system while the technique of Long Dialogues is new. The content of this part show a preference to more authentic materials than before, such as conversations, academic lectures, and TV programs, etc. At the same time, the proportion of reading is decreased from 40% to 35%. The item of comprehending reading passages is in concord with the old test, whereas Banked Cloze as well as Skimming and Scanning are new. In addition, there are two more newly-introduced types of question, Translation from Chinese to English and Error Correction. Apart from these changes, an oral test is also added, which is held months later as an independent part that students can take, if they wish to, based on their performance on the written test.

The time allotment for the new test is 125 minutes. Students are supposed to do the writing task in the first place, followed by Skimming and Scanning task, after which they have to

hand in their answers of these two parts. Then they do the rest of the test in this order – Listening Comprehension, Reading in Depth, Cloze and Translation from Chinese to English.

3.1.2 Score report

In the old CET-4 test, a single score, or a composite score (Bachman & Palmer 1996:223) was reported, with 100 as the full mark and 60 as pass, and two types of certificates were awarded: fair (60 to 84) and excellent (85 and above). The new scoring system, starting from June 2005, takes scores ranging from 290 to 710, with 710 as the full mark. No criteria are set as pass or excellence and no certificate with fairness or excellence is awarded. Instead, each candidate receives a profile of scores which contains a total score as well as four other scores which correspond to the different areas of language ability intended to be measured, i.e. Listening, Reading, Comprehensive skills and Writing. The reform on the score report is taken to facilitate teaching, avoiding unreasonable comparison between different colleges, and facilitate learning, keeping students out of the pressure from the job market and graduation (Guo 2006a).

3.2 Reliability

The degree of the reliability of the CET-4 test can be measured in several ways. To begin with, there are two components of test reliability: the performance of the candidates from occasion to occasion, and the reliability of the scoring (Hughes 2003:44). Let us look at the data provided by Yang (2003) in his paper that reviews the CET-4 test in 15 years since establishment. The scores of students in the year 1996, 1997 and 1998 from two sample universities were collected. In School A, the percentages of students who passed the test in three years are 86.50%, 93.21% and 94.05%, respectively. In School B, the percentages are 22.70%, 22.75% and 22.94%, respectively. From the data above, we find that the performance of students in each school generally remained stable throughout three years and the reliability estimates were well within the desirable range and substantial.

Now let us look at the data provided by the National CET-4 and CET-6 Commission

concerning the percentage of students who passed the new CET-4 in June 2006, December 2006 and June 2007, from universities and colleges in China and those in Beijing municipal city. The national percentages of students in these three tests are 19.20%, 28.20% and 26.40%, respectively. The percentages for students in Beijing are 32%, 43.10% and 41.20%, respectively. The June 2006 test is the first time the new system was operated nation-wide, and this is possibly the reason of relatively low percentages shown above. However, we could still conclude that students' performance remained relatively stable and the reliability estimates were within the range. Thus, both the old and new CET-4 tests are said to be objective and highly reliable.

Secondly, due to the relationship between validity and reliability, test developers tend to set the minimum acceptable level of reliability as high as possible. As has been pointed out, candidates should not be allowed too much freedom on choosing test items, in order to achieve reliability. Bachman and Palmer (1996:135) provide two criteria to evaluate in this respect: one is "the way the construct has been defined", and the other is "the nature of the test tasks". That is to say, only when the test focuses on a relatively narrow range of components of language ability with relatively uniformed test tasks, could the test achieve higher levels of reliability. In the writing part of the old and new CET-4 tests, a controlled composition is adopted. What the students should write is clearly specified and illustrated by different ways such as an outline, charts and tables, key words, or pictures. Controlled composition is, therefore, seen as beneficial to the improvement of scoring consistency.

Thirdly, clear instructions and unambiguous test items can contribute to the reliability of a test. Both tests have done well in this regard. For example, the outlines in the writing part of both tests are given in Chinese in order to avoid vagueness and to keep students from copying the original English sentences. Apart from this, one kind of instruction is given in Chinese, too. That is, there are instructions which direct test-takers to write their responses on the corresponding answer sheets. This also works well against vagueness. The other instructions, which are given in English, are not only provided with a heading, but accompanied by specific directions explaining what candidates are expected to do (see Figure 8). This insures

that there will not be misunderstanding by the candidates or an unanticipated answer. In addition to the instructions, most items in both the old and new tests are in formats that are familiar to candidates, who are likely to have done many simulated and previous papers as practice before they take the actual test.

Fourthly, when we make a comparison of the objective and subjective items and their scoring between the old and new CET-4 sample tests, we find that the old test, out of 91 items, includes 90 multiple-choice questions, constituting 98% percent of the total score. In contrast, the new system contains 80 multiple choice items out of 92 items, accounting for 76% of the total score. As far as testing technique is concerned, the old system is of higher reliability than the new one.

The old CET-4 test	**Part II Reading Comprehension (35 minutes)** ***Directions:*** *There are 4 passages in this part. Each passage is followed by some questions or unfinished statements. For each of them there are four choices marked A), B), C) and D). You should decide on the best choice and mark the corresponding letter on the Answer Sheet with a single line through the centre.*
The new CET-4 test	**Part II Reading Comprehension (Skimming and Scanning) (15 minutes)** **Directions:** In this part, you will have 15 minutes to go over the passage quickly and answer the questions on **Answer Sheet 1.** For questions 1-7, mark Y (for YES) if the statement agrees with the information given in the passage; N (for NO) if the statement contradicts the information given in the passage; NG (for NOT GIVEN) if the information is not given in the passage. For questions 8-10, complete the sentences with the information given in the passage.

Figure 8: Some of the instructions in the two sample tests: the old and new tests are very similar with this regard.

Finally, although there are major changes in the test contents, the overall length of the new system does not make much difference. However, the part of Listening Comprehension has been significantly lengthened, from two question types to four, resulting in an increase in the number of test items, score proportion and time. As argued by Hughes (2003:44), the degree of reliability varies in proportion to the number of items, if other things are equal. That is, the more items are included in a test, the more reliable the test becomes. The new test is thus

considered to have a higher reliability in testing the listening ability of candidates than the old system. The reading part is a different case. It removes two passages which adopt the multiple choice format, and introduces a Skimming and Scanning passage and a Banked Cloze passage. As far as the length is concerned, this does not make much difference. However, it facilitates a higher degree of validity, thanks to the various testing techniques involved. As has been pointed out in section 2.3.2, a necessary condition as it is, reliability is not sufficient for construct validity and test usefulness (Bachman & Palmer 1996:23). A multiple-choice test might yield very consistent or reliable scores, but this would not sufficiently justify using this test in measuring the overall listening or reading competences.

From this brief comparison between the two tests, it is difficult to draw a conclusion as to which test is more reliable. In addition, to assess the reliability of a test, a variety of other factors need to be taken into consideration, such as the representativeness of test items, scoring procedures and so on. One effective way is to quantify the reliability in the form of a reliability coefficient. As this is beyond the scope of this essay, it will not be further discussed here.

3.3 Validity

In order to assess the validity of a test, two forms of evidence are needed (Hughes 2003:26). The first form relates to the content of the test. It should be demonstrated that the test content constitutes a representative sample of the language skills that are to be tested. The second form, namely, criterion-related validity, is related with the degree of the test scores matching the candidates' ability. The National CET-4 and CET-6 Commission in China spent three years on the assessment of the validity in a full-scale manner and presented the results in *Validation of the National College English Test*, published in 1998 (Yang & Weir). This essay, hence, will only focus on investigating the contents of the two sample tests and make a comparison between them in this respect.

The new specifications provide a more explicit definition of constructs to be tested, than the old version. That is, the new specifications define clearly the four abilities to be measured in

terms of their contents and methods. For example, in testing the listening ability, the new version stipulates the skills that are to be tested and provides detailed information about testing techniques included in the listening part, not to mention the accents and reading speed that are used. In contrast, the old specifications contained only general information, which might cause misunderstandings during the course of test development, and result in the invalidity of the test. As has been pointed out in Section 2.3.1, explicit specifications which take account of all that is known about the constructs, contribute to the validity of a test.

3.3.1 Listening

The old Listening Comprehension part comprises only two test types, namely, Short Conversations and Listening Passages / Compound Dictation, which candidates are expected to finish within 20 minutes. Generally, the first section involves either academic conversations, such as one between a professor and a student about a course assignment, or functional conversations of daily life, such as making a holiday plan between friends. Test-takers have to not only distinguish the pronunciations and intonations so as to understand the literal meaning of speakers, but to listen between the lines as well, and give a correct response out of four choices. The part of Listening Passages is somewhat easier since there are more context cues. Yet still, test-takers must comprehend the gist of the passage and identify the specific information while working out a correct response from four choices. The section of Compound Dictation is seen as the most difficult item to test-takers since it not only tests how they receive information, but how well they produce what they have heard in the target language.

The dominant discrepancy between the old and the new tests in the listening part is the inclusion of Long Conversations and Compound Dictation. The latter is used as a regular item in line with Listening Passages, and candidates are required to complete the four sections within 35 minutes. This change increases the number of questions considerably, and puts candidates under greater stress Thus, this significant change is thought to measure candidates' listening ability more precisely than the old test. As far as the testing technique is concerned, while the old system mainly uses a multiple-choice format, the new test adopts varied formats.

When referring to multiple-choice questions, Bachman and Palmer argue that dichotomous scores are not effective indicators of proficiency levels (1996:150). Furthermore, this discrete response format promotes testing of the formal linguistic system (McNamara 2000:14), where "localized grammatical characteristics" rather than broader, global discourse skills are the focus (Buck 2001:123). That is, multiple-choice is considered to hamper the score generalizability for the domain of generalization, the university environment. Conversely, the listening part in the new test, taking its format from the TOEFL test, is more synonymous with the trend towards pragmatic and integrative testing, which is driven by today's communicative trend and subsequently productive language use focus (McNamara 2000:14). In other words, the new CET-4 test uses a variety of test tasks which: 1) engage the students in different areas of listening language ability; and 2) correspond better to the university environment than any single response format. This will be further clarified in the discussion of authenticity in section 3.4.

3.3.2 Reading

As far as the reading part is concerned, after a close look into the January 2002 test paper, we find that it contains four pieces of comprehension passages, covering a considerable range of topics involving technology, biology, and social sciences. The candidates are allowed 35 minutes in reading the passages as well as completing the comprehension tasks. Including the questions, the whole reading part constitutes 1,997 words. This means that the test-takers need to read at a rate of approximately 57 words per minute. While in the new test, candidates have to finish the comprehension of 1,228 words within 20 minutes, requiring them to read at a rate of 61.4 words. Either 57 or 61.4 is a high demand for EFL learners at this level. Candidates also need to prove their abilities in language knowledge as well as background knowledge.

What is more, the reading part not only questions the related information but also the implied meaning and even the specific meaning of a certain word. To do these, candidates need the skills of reading both extensively and intensively. If “the purpose, events, skills, functions, levels are carried out as what they are expected to” (Carroll 1980:67), the construct validation is fully displayed in the reading part.

One of the major differences between the old and new CET-4 tests concerning the content lies in the inclusion of Skimming and Scanning. From the sample test, we see that Skimming and Scanning is fundamentally in line with the new specifications, which state that candidates should be able to "read, at a rate of 100 words per minute, long passages of less difficult, using 1) skimming skill to obtain the main idea and 2) scanning skill to obtain specific information". This is a significant improvement in that the new test includes a broader coverage of reading abilities to be tested.

3.3.3 Vocabulary and grammar

The new test removes the part of Vocabulary and Structure, but introduces a new type of question, namely, Banked Cloze (see Figure 9).

<table>
<tr><td>The old CET-4 test

Part III Vocabulary and Structure</td><td>41. By the time you get to New York, I _____ for London.
A) would be leaving B) am leaving
C) have already left D) shall have left

48. In the Chinese household, grandparents and other relatives play _____ roles in raising children.
A) incapable B) indispensable C) insensible D) infinite</td></tr>
<tr><td>The new CET-4 test

Part IV Reading in Depth

Section A</td><td>When Roberto Feliz came to the USA from the Dominican Republic, he knew only a few words of English .Education soon became a __47__. "I couldn't understand anything," he said. He __48__ from his teachers, came home in tears, and thought about dropping out.
Then Mrs. Malave, a bilingual educator, began to work with him while teaching him math and science in his __49__ Spanish. "She helped me stay smart while teaching me English ,"he said .Given the chance to demonstrate his ability, he __50__ confidence and began to succeed in school.
[…]
<table><tr><td>A) wonder</td><td>I) hid</td></tr><tr><td>B) acquired</td><td>J) prominent</td></tr><tr><td>C) consistently</td><td>K) decent</td></tr><tr><td>D) regained</td><td>L) countless</td></tr><tr><td>E) nightmare</td><td>M) recalled</td></tr><tr><td>F) native</td><td>N) breakthrough</td></tr><tr><td>G) acceptance</td><td>O) automatically</td></tr><tr><td>H) effective</td><td></td></tr></table></td></tr>
</table>

Figure 9: Testing of vocabulary and grammar in the old and new CET-4 test

While the old CET-4 test assesses grammatical and lexical knowledge in isolation with multiple choice questions as the technique, the new system presents adequate contextual cues, requiring candidates to actually “use words, phrases and grammatical structures” (Specifications for the CET-4 [Revised Edition 2006]). Again, this technique is in better accord with the trend towards pragmatic and integrative testing. Further discussion concerning this will be conducted in section 3.7.

3.3.4 Score report

An important component of test validity is the scoring system. In addition to the scoring of test items and score interpretation (which are beyond the scope of this essay and will not be elaborated here), the way scores are derived can make a difference.

The former way, where a single score was reported, could inform candidates of their overall language ability. Also, a compensatory composite score could be derived (Bachman & Palmer 1996:224). As is mentioned in section 2.3, a candidate with high levels in some of the areas of language ability can use these high levels of ability to compensate for low levels of ability in other components, and consequently arrive at a balance of the high and low scores. However, this method cannot precisely inform a candidate of specific language ability, and the scores cannot be used for his (or her) specific purposes.

The new way of reporting scores, where a profile of scores that contains four scores which correspond to the different areas of language ability as well as the sum of these component scores, can work well to attain all the above-mentioned purposes. For example, it can effectively inform a candidate of his strengths and weaknesses of his language abilities before he proceeds to make plans for improvement. Another example would be that this profile of scores can be used to make decisions about a position which requires an applicant to have good competence in listening and speaking, while having an average level of reading and writing. To sum up, the new method of reporting scores provides a more valid and reliable way of assessing language abilities.

3.3.5 Summary

The major changes in the test specifications, contents, as well as in the way of reporting scores can well lead us to the impression that the new test is of higher validity than the old one. Firstly, the new specifications state more explicitly what to be measured and how to do it. Secondly, the new test covers a wider range of skills to be tested by adopting varied testing techniques in the Listening Comprehension and Reading Comprehension parts. Moreover, vocabulary and grammatical structures are not tested in isolation, but are measured within contexts. Finally, the way of reporting scores enhances the validity and reliability of the new test.

3.4 Authenticity and interactiveness

Authenticity must be assessed from a number of perspectives (Bachman & Palmer 1996:25), such as input, language, task and content. In this section, the focus is on the content authenticity.

3.4.1 Identifying the TLU domain

The TLU domain is a set of settings and tasks that the students are likely to encounter and that require target language use. The CET-4 claims to measure college students' overall capacity in English, especially their listening and speaking competences, and improve both their oral and written communicative abilities for their future careers and social activities. Moreover, before they use English in their future careers, students are initially exposed to English in an academic situation. Thus the TLU domain comprises both academic and non-academic settings.

In Chinese university contexts, students use English in academic situations, though the amount is rather limited. For example, most of the activities in English classes are carried out in English, though not necessarily by native speakers. In the English-related courses, students listen to instructions by the teacher, ask questions, give responses, take notes, have discussions with both the teacher and classmates, write essays and reports, and take tests.

They have to read classroom-related materials and do after-class assignments using their English textbooks. However, they seldom use English for communicative purposes out of class. They do not have much chance of engaging in social activities with foreigners, either.

In their future careers, they might use English in a specific career-related setting, such as in an office where they have to read and write emails, read various instructions and tables, and sometimes communicate orally with clients.

3.4.2 Authenticity in Listening, Reading and Writing

In this section, degrees of authenticity of different parts will be measured. First and foremost, the respective features of the various parts of the two tests will be presented adjacently in a table format (see Table 3). Then sample questions for both tests will be looked at by analyzing their formats (method and construct), and contents (situations), after which the test task's levels of authenticity will be critiqued and discussed.

3.4.2.1 Listening

As far as listening ability is concerned, the materials selected by the old CET-4 test contain "daily-life conversations with less complex topics and sentence structures, as well as stories, talks and narratives of general topics that most students are familiar with" (*Specifications for the CET-4 2005).* The new CET-4 test requires students to be able to perform both academic and non-academic listening related to language functions as follows: "to listen to lectures given in English, daily life conversations, lectures of general topics and to understand English programs conducted with a mediate-speed of 130 words per minute, obtaining the gist of such talks" (*Specifications for the CET-4 [revised edition 2006]).*

Table 3 shows whether the old and the new tests have included elements of authenticity or not. Test questions have been taken out from two sample tests, categorized and analyzed.

Table 3: Authenticity of the Listening Comprehension part in the old and new CET-4 (A 'O' represents that there is a high correspondence between the test task and the TLU domain, while a "X" represents that there is a lack of correspondence. A "△" stands for an uncertainty of the degree of correspondence).

Category	**The old test**		**The new test**	
	Test	**TLU**	**Test**	**TLU**
Situation	*Conversations:* 1) About sending emails between colleagues 2) About the close time of the library between friends 3) About the operation of a machine between friends 4) Apologizing between friends 5) Saying goodbye to a friend 6) About future plan between friends 7) About current position between friends 8) About tickets between friends 9) About a lecture between friends 10) About a job interview between friends *Passages:* 1) A story about a man and his bird 2) A description of British legal system 3) A description of the London driver	△ X X X X X X X X X O O O	*Short conversations:* 1) About borrowing and returning things between friends 2) Giving suggestions between friends 3) Talking about a place with friends 4) At an office between an employee and a boss 5) Talking about a lecture between friends/classmates 6) Describing the campus 7) Talking about a newspaper article between husband and wife 8) Talking about newspaper sections between friends *Long conversations:* 1) At an office between colleagues about work proposals 2) Choosing an essay topic *Passages*: 1) An article about future transportation 2) A description of a western custom 3) An introduction of the Cambridge University *Compound dictation:* 1) A news report about Russia's entry into WTO	X X X △ O X X X △ O O O O O
Format	1) Multiple choice questions (reading of choices involved)	X	1) Multiple choice questions (reading of choices involved) 2) Gap-filling (writing and reading involved)	X O

Table 3: Authenticity of the Listening Comprehension part in the old and new CET-4 (Continued)

Category	The old test		The new test	
	Test	TLU	Test	TLU
Strategic competence	1) Summarizing main ideas 2) Locate specific information 3) Making inferences 4) Understanding special expressions	O O O O	1) Summarizing main ideas 2) Locate specific information 3) Making inferences 4) Note-taking 5) Understanding special expressions	O O O O O

The listening parts in both tests include lectures that students might encounter in an academic setting (e.g. the Passage sections). Although the likeliness of daily-life conversations between friends is relatively low, it simulates the speaking environment in English-speaking countries, and adds some idiomatic expressions common to spoken English to attain the features of the target language use. In addition, the new specifications claim that the CET-4 test measures college students' overall ability in English, especially their listening and speaking competences, and prepare them for both oral and written communication in their future careers and social activities and that the score obtained from the test is used to ascertain whether a student can operate competently in jobs and daily life. These requirements are mapped in the two tests, too. Therefore, we could say that this section provides authentic materials to a limited extent.

However, when comparing the old test column with the new test column as shown in Table 3, common features for both can be clearly seen. For the new test, there are more circles (O) in its respective TLU column than for the old test. This is primarily attributed to the increased amount of contents and testing techniques. This gives rise to the argument that the listening part in the new CET-4 has higher authenticity than the old CET-4.

There is an obvious problem with the content, though. Regardless of the functional conversations such as borrowing things in daily life, the EFL environment for college students in China is mainly the campus. The TLU required is that of academic situations, not borrowing things from a friend. The conversations of the Listening Comprehension part simulate those in a real-life situation, which is less likely to happen. On the other hand, few

conversations in academic settings are include which is where students are supposed to use English most often.

3.4.2.2 Reading

The EFL environment for college students in China is obviously the English-related classroom, where they encounter articles of various topics in their textbooks and other materials, including the ones specified in the specifications. Apart from that, the new specifications emphasize that students have to read materials relevant to their specific careers in the future. That is, they have to read memos, notices, emails, instructions, tables, letters, reports, schedules and so on. These should all be included in the reading materials.

Table 4 shows a comparison of the old and new tests in terms of the authenticity of the reading part.

Table 4: Authenticity of the reading part in the old and new CET-4

Category	**The old test**		**The new test**	
	Test	**TLU**	**Test**	**TLU**
Situation	1) A report about the future of automobiles	O	*Fast Reading*	
	2) A description about fox hunting	O	1) About landfills	O
	3) A report about the aging issue in America	O	*Reading in Depth*	
	4) A report about moral decline	O	1) A biographic story concerning bilingual education	O
			2) An argumentation about sports	O
			3) A report about environment	O
Format	1) Multiple choice questions	X	1) True or false statements	X
			2) Multiple choice questions	X
			3) Gap filling	X
Strategic competence	1) Summarizing main ideas	O	1) Skimming and scanning	O
	2) Searching specific information	O	2) Summarizing main ideas	O
	3) Making inferences	O	3) Searching specific information	O
	4) Distinguishing opinions and attitudes	O	4) Making inferences	O
	5) Comprehending word meaning	O	5) Distinguishing opinions and attitudes	O
			6) Comprehending word meaning	O

As can been seen, the contents included in both the old and new CET-4 tests bear authenticity to a large extent. The passages in both tests contain a wide range of topics, covering the society, environmental issues and education, which are likely to be encountered in English textbooks as well as in newspapers. The strategic competences required of the students are highly authentic as well. Nevertheless, the new test is of higher authenticity than the old test in that the former contains more varied test formats. In the TLU domain, people are usually expected to read at a fast speed searching for specific information that they need. The section of Skimming and Scanning thus serves this purpose very well.

There is a problem with the new test, though. As is stipulated by the specifications, students should be able to read “practical reading materials in their future careers and daily life”. However, we cannot find a single passage of this kind in the new test. In fact, practical reading materials are usually a crucial component in students’ future careers. The exclusion of practical reading materials reveals that there is a lack of correspondence between the test tasks and the specifications, and between the test tasks and the TLU domain, thus lowering the degree of the test authenticity.

3.4.2.3 Writing

The TLU domain of reading tasks applies to the writing task, too. It includes both academic and non-academic settings, where students have to write academic essays on the one hand, and prose for practical purposes on the other. These should be taken into consideration when designing a test. A similar comparison of the old and the new tests with this respect is administered in Table 5.

Regardless of the improvement in the specifications, the new CET-4 test remains almost unchanged in terms of format and strategic competences. Both tests reflect in detail the writing competence required of students in the academic setting, although it is not specified in the old specifications. The formats contain elements of practical writing, i.e. a letter and a speech. Therefore, it is argued that both the old and the new tests are similarly authentic to a certain extent.

Table 5: Authenticity of the Writing part in the old and new CET-4

Category	The old test		The new test	
	Test	**TLU**	**Test**	**TLU**
Situation	1) Writing a letter to the university principal about the school canteen	X	1) Writing a campaign speech as a candidate for the president of the Student Union	X
Format	Writing a letter	O	Writing a speech	O
Strategic competence	1) Giving personal opinions 2) Describing things 3) Giving suggestions	O O O	1) Giving personal information, i.e. personality, hobbies, etc. 2) Describing future plans	O O

The problem, however, is salient. Writing for practical purposes such as memos, notices, invitations, emails, etc., which are usually used in the non-academic setting, are excluded in the writing part due to the fact that these tasks produce only limited output and cannot represent the overall writing competence of students. But contradictorily enough, they are explicitly stated in the new specifications and more importantly, are the most frequently used writings that students might encounter in their future careers.

3.4.3 Summary

By comparison of three parts between the old and the new CET-4 tests, we find that the new CET-4 test is of higher authenticity than the old test as far as contents are concerned. This is mainly attributed to the addition of contents and adoption of varied testing techniques, such as Long Conversations in the listening part and the Skimming and Scanning in the reading part. However the problem is apparent, too. The new test, the reading and writing parts in particular, does not include contents that students might encounter in the non-academic setting. Although the listening part contains elements of the real-life setting, the Conversation section lacks focus on the academic setting.

3.5 Impact

The discussion on impact will basically be elicited from the results of three previous empirical studies. The first is a case study administered after the reform of the CET-4 test, by the faculty

in a university in South China, where the English learning processes of six Chinese students within the first two academic years in university were recorded and examined in attempt to investigate the power of the CET-4 over what took place outside the classroom (Zhan n.d.). The second study looks into the washback effect of the CET-4 on teachers (Chen 2007). The third study was conducted shortly after the first reformed CET-4 test, by faculty from an institute in central China, in the form of questionnaires, aiming at validity of the CET-4 and its impact on teachers, learners as well as employers (Wang et al. 2005). Following this, there will be a brief comparison between the old and new CET-4 tests in this respect.

The validity of the three studies has been discussed in Section 1.3, so it will not be mentioned again here.

3.5.1 Impact on learners

Three results are elicited from the first study: 1) the CET-4 test exerts some influence on students' learning behaviors outside the classroom; 2) The extent of the CET4 washback on learning becomes increasingly greater when the exam approaches nearer; 3) Even if the test format and content are changed, this will not automatically bring about the intended washback on all learners since there are other factors involved in the mechanism of washback.

According to Zhan (n.d.), a student's learning processes is composed of three stages, which are 1) the period of accommodation to College English learning, 2) the period of normal College English learning and 3) the period of preparation for the CET-4. The impact of the CET-4 on learners becomes increasingly prominent as the test date draws near. For example, students tend to show concern with the test and do relevant exercises occasionally during the first and second periods, but this is not common among students (Zhan n.d.). However, as the test approaches, students focus increasingly on it and spend more time on the preparation. Some students stop their regular study routines and exert all their effort on the test preparation. Wang et al. (2005) discover that some students even take up time that should be used for other courses in preparing for the test, such as memorizing vocabulary lists, doing past papers, and attending refresher courses. In a word, the CET-4 test motivates the students in learning to a

large extent.

Since it is revealed that the CET-4 test has influence over students' learning behaviors (Zhan n.d.), it may be assumed that students would subsequently make modifications in their learning contents and styles as a response to the change of the test. For example, the increased percentage in the listening part would cause students to do listening practices voluntarily, which most of them have found difficult and used to feel reluctant to do. On the contrary, though the new system includes an oral test, which only applies to the minority of candidates who have performed with distinction on the written test, and more importantly, whose grade is not included in the total score, the majority of students are assumed not to spend much time in this respect, unless out of his personal interest. In this stage of quantitative learning, both learning contents and styles would be adjusted mainly as required by the test. As can be seen from the survey, students, on the whole, tend to adopt an exam-oriented approach when they begin the preparation, while as they admit the importance of communicative function of English, they give priority to the written test.

The CET-4 test initiates students in learning on the one hand, whereas on the other hand, it puts students under great pressure and inevitably meets with resistance from students, due to its high stakes and its power over students' future. According to Wang et al. (2005), 39.9% students feel it unworthy to spend so much time on English that they neglect other courses. 19% students consider the CET-4 test as "a nightmare". Therefore, research has to be done as to reduce the negative washback effect of the test to the minimum while maximizing its positive effect.

3.5.2 Impact on teachers

Wang et al. (2005) discovered the ambiguous feelings teachers have to the CET-4 test. On the one hand, most English teachers see the CET-4 test as a powerful force in promoting the English learning context in China. This is inspiring for English teachers in China, since the importance of the test influences the status of their job which is thought to play a significant role in the educational system. Nevertheless, the study also reveals that 37.7% of the teachers

believe that the test itself has become a baton to lead their instruction. 62.3% of the teachers doubt the test's power in actually improving students' English knowledge and 79.1% of them do not believe the test could enhance teaching in general.

Such feedback from teachers has confirmed the great influence, both positive and possibly negative, of the CET-4 test on teachers.

The CET-4 has an influential impact on teachers' curricular planning and instruction. (Chen 2007). Since it is inevitable to "teaching to the test", teachers tend to modify the teaching contents in line with the requirements of the CET-4 test, especially as the test draws nearer. For example, the listening part took up only 20% in the old test while reading accounted for 40%. As a result, teachers gave more instruction on improving the reading ability than on listening. That is possibly the justification to the fact that most teachers in Wang et al. (2005) study show doubts for the CET-4 test, which confirms that the old CET-4 test yielded negative washback effect in the present EFL context. The low degree in authenticity of the old test as reflected in teaching, instructions were mainly based on grammar, ignoring the communicative function of language. It is found that teachers and students spent a tremendous amount of time preparing for the CET-4 test so that training for the test dominated the classroom.

Unlike the old sysem, the new test, which is designed on the basis of the communication-oriented specifications, gives equal weight to listening and reading. As a result, the classroom instructions on listening have been increased. Chen (2007) points out in his study that "due to the new testing objectives，teachers [were] motivated to integrate listening and speaking into their teaching，especially listening，instead of only teaching reading and writing as the previous case".

However, it should be admitted that such a change might not be simply due to the new CET-4 specifications but to the change of teaching materials as well. The designated teaching material package, compiled and issued by the Ministry of Education，included textbook(s), teachers' manual(s), and tape(s). And listening has earned more concern in the new teaching

materials than in the old ones.

An interesting finding in the study by Wang et al. (2005) is that 70% teachers do not think of abandoning this national test as a good idea, regardless of the almost unanimously negative opinions toward the test. This is probably because the CET-4 test as a whole has brought a positive washback effect on the status of English as well as on the English teacher profession, yet at the same time there is an urgent need for the innovation of the old system.

3.5.3 Impact on society and educational system

Bachman and Palmer demonstrate that when the impact of a particular test use is being assessed, what should be taken into consideration are the characteristics of the testing situation (such as the TLU domain and construct definition) in terms of the values and goals of the test-takers and of the educational system and society (1996:35). For example, the use of multiple choice questions, which is the dominant type of test tasks in the old CET-4 test, was thought to influence the whole language teaching practice and language programs of educational system throughout Chinese universities and colleges. An idea would be formed by teachers, students as well as the test users that language abilities are assessed and improved mainly by doing multiple choice questions, and such an idea will consequently give rise to the dominant role of multiple choice items in teaching practices.

Shohamy (1998) argues that "[t]ests are most powerful as they are often the single indicators for determining the future of individuals. As criteria for acceptance and rejection, they dominate other educational devices such as curriculum, textbook and teaching." This is particularly true for the CET tests. As has been mentioned in the beginning of this section, employers use it as a criterion to assess whether an applicant is qualified for the job in terms of his English proficiency. To a certain extent, the CET-4 test has "imposed implicit ideas about success, knowledge, and ability" of an individual (Noam 1996, as cited by Shohamy 1998). This is confirmed by the fact that an applicant with a CET certificate in China has better chances of getting a satisfactory job than one without a CET certificate.

As a large-scale nation-wide test, the reform of the CET-4 test will inevitably bring about significant impact on the society and the educational system, especially since it includes a new oral test. Due to the short time since its introduction, time is still needed for the washback effect of this change to show.

3.6 Practicality

Both the old and new CET-4 tests are high in practicality. The National CET-4 and CET-6 Commission, under the supervision of Ministry of Education, is responsible for the CET tests both academically and organizationally. The commission is composed of professors and experts from several top universities in the country. It has two consultants, one commission chairman, several vice-chairmen as well as some other professionals and consultants.

The National CET-4 and CET-6 Commission is responsible for the organization of the CET-4 test. In addition, part of the administration is conducted by the Examination Center of the Ministry of Education.

As a nation-wide test that has been administered for over two decades, the administrative work is complex yet well managed. The educational institutions concerned at all levels work coordinately to ensure the smooth operation of the test, in terms of test development, administration and scoring.

3.7 Testing grammar and vocabulary in the new CET-4 test

In the framework of the language structure put forward by Bachman and Palmer, learners' language ability is defined as involving two components: language knowledge and strategic competence/metacognitive strategy (1996: 68-75). That is, learners need to know the vocabulary, grammar, sound system as well as to use coherent sentences in a certain language setting to achieve the communicative goals of language users. The CET-4 test, a way to demonstrate candidates' achievement in English, should determine whether they could apply the knowledge and skills in their future careers and real-life communication, that is, to assess their performance in this language. However, we do not expect a test to measure all the

aspects of language in each section. Thus, the samples should be as representative as possible.

In Bachman and Palmer's framework, grammatical knowledge includes knowledge of vocabulary, syntax, phonology, and graphology (1996:70). In the new CET-4 sample test, knowledge of vocabulary seems to be tested in all the parts, which proves the common sense that words are basic building blocks of language.

3.7.1 Listening

In Listening Comprehension, testing vocabulary is not limited to single words. There are many compound words, phrases and even idiomatic expressions and slang. For example, in the sample test, there are many phrasal verbs like *go over*, *come up with*, *keep ahead of competitors* and *cut down* in the dialogues between two speakers. Colloquial expressions like *get started with* and *I've got to do something* are found, too. Occasionally, there is slang like *what's the picture*? Since most of the dialogues are selected from daily-life conversations in English-speaking countries, some phrases and sentences could cause difficulty for EFL test-takers since it is difficult to work out the meaning by the surface meaning of the words. Moreover, both the conversation and the choices are highly demanding on grammar to require the test-takers to give a definite response in 15 seconds. For example, question number 15 presents three different tenses, namely, the simple present, the present continuous and the future tense as well as several sentence patterns, i.e. yes-no questions, interrogative questions, the attributive clause and the objective clause (see Figure 10).

(from the tape script)

15. M: Do you want to go to the lecture this weekend? I hear the guy who's going to deliver the lecture spent a year living in the rain forest.
W: Great! I'm doing a report on the rain forest. Maybe I can get some new information to add to it.
Q: What does the woman mean?

(from the test paper)

15. A) She knows the guy who will give the lecture.
B) She thinks the lecture might be informative.
C) She wants to add something to her lecture.
D) She'll finish her report this weekend.

Figure 10: Question number 15 in Listening Comprehension of the new CET-4 test

From this short dialogue, we notice that usually the first person presents the content or the background of their conversation and the second person gives the hint to the answer of the question. Summing up the questions from the first thirty short dialogues, we get the following results (see Table 6):

Table 6: 15 Questions of dialogues (short and long) in the new CET-4 sample test

Typical questions	**Percentage**
What is/are the man/woman/speakers doing/talking about?	27%
What does the man/woman suggest?	20%
What does the man/woman say about…?	13%
What does the man/woman mean?	7%
What is the probable relationship between the two speakers?	7%
What can we learn / be inferred?	3%
Others	13%

From the types of questions, it is not very difficult to see that answering these questions in "Listening Comprehension" needs both fluency and consolidated grammatical knowledge. The Listening Comprehension test is much more a combination of testing on both vocabulary and syntax.

3.7.2 Reading

The communicative philosophy of a reading test is to test "in what situations do we read which texts for which purposes" (Wijgh 1995). The old CET-4 test had vocabulary items, which were selective and context-independent multiple-choice items presenting words in isolation. They were criticized since students simply spent time unproductively memorizing long list of words together with synonyms or definitions. Apart from this, the testing on vocabulary items was a salient feature in the Reading Comprehension subtest, too. The reading part usually included 1-2 questions out of 20 about words, such as Question 31 in the January 2002 test paper (see figure 11).

31. "…Old is suddenly in" (Line 1, Para. 1) most probably means "_____".
    A) America has suddenly become a nation of old people

B) gerontology has suddenly become popular
C) more elderly professors are found on American campuses
D) American colleges have realized the need of enrolling older students

Figure 11: Question number 31 in the January 2002 test paper

These questions were always demonstrated in several fixed ways: “The word ‘appraisal’ in line 2, para. 1 is closest in meaning to…”; “the phrase ‘out of sight, out of mind’ in line 2-3, para. 3 means…”. Although some questions concerned much of the word meaning in the context, other word questions seemed to assess the range of candidates’ vocabulary. And sometimes without referring back to the contents, test-takers could still get the answer if they simply know the meaning of words. These vocabulary items in the reading test could be categorized into the relative independent group, despite the manner in which they were presented.

The new CET-4 test removes the part of Vocabulary and Structure and at the same time has a subtest in the reading part. The subtest Banked Cloze is a short passage where there are 10 blanks that require candidates to fill in with the appropriate words from 15 choices (see Figure 9). This technique measures vocabulary in line with its embedded, comprehensive and context-dependent nature. A candidate has to consider the word in terms of its meaning, collocations, relationship within the context and so on.

3.7.3 Writing

The writing part checks not only the written form of the words but also the function and collocations of their grammatical usage. Cumming and Mellow (1995) define a general ESL composition profile, which is “vocabulary (range, choice, usage, word form mastery, register), language use (complex constructions, errors of agreement, tense, number, word order/function, articles, pronouns, prepositions) and mechanics (spelling, punctuation, capitalization, paragraphing)”. In the CET-4 test, candidates need to finish a composition in 30 minutes, which is constituted by at least 120 words. However, the weakness in this part is in its limited styles of writing. Like the topics in the sample tests, most of the writing styles involve academic writing showing one’s preference or opinions. Although the writing part is not the

specific section to test vocabulary and grammatical knowledge, it is still a question whether the sample chosen in the test is truly the representative of the communicative competence.

3.7.4 Cloze

In language proficiency testing, Read (1993:357) recommends that words need to be understood in connected written or spoken discourse rather than just isolated items, which is very important to the EFL learners. In a grammar test, the content should be defined more broadly than “syntax and morphology” and includes textual competence as well (Rea-Dickins 1997: 92).

Although a cloze passage is usually considered to be testing the overall ability (Hughes 2003:186-194), and to respond to the deletions in a passage successfully needs more than grammatical ability, it is generally accepted that most of items in a cloze passage are inclined to test either grammatical structures or vocabulary. Take the new CET-4 sample test as an example. Out of 20 items, 18 items test the vocabulary and 2 items test grammatical knowledge of a candidate. Therefore, it is admitted that to perform successfully on a cloze test requires integrated skills. But more importantly, a good master of vocabulary and grammatical knowledge is the best guarantee.

However, the cloze procedure is not without problems. The first problem is with its content validity and reliability. Hughes argues that different passages give different results, as does the deletion of different sets of words in the same passage (2003:189). Consequently, there is a need for careful selection of texts and some pre-testing. Secondly, as the cloze test in the CET-4 adopts the format of multiple-choice questions, it is usually difficult to write successful distractions. For example, in Question number 68 of the sample test, candidates might easily exclude choices B) and C) once they realize they should have a word that collocates “your money” (see Figure 12). Question number 77 has the same problem since choices C) and D) can be identified as unqualified without much difficulty. The third problem resides in the selection of items to be tested. Once a passage has been chosen, we are limited as to what features of language can be tested. Sometimes, there might be an unbalanced coverage of the

features. The cloze test in the new CET-4 sample test paper, for example, includes 18 items of vocabulary and only 2 items of grammar. This inequality in distribution makes the cloze procedure, instead of testing integrated ability, a vocabulary test that resembles in nature the Banked Cloze subsection in Reading Comprehension, but much easier.

Wise buying is a positive way in which you can make your money go further. The __67 (way)__ you go about purchasing an article or a service can actually __68__ your money or can add __69 (to)__ the cost. […] The cost of the electricity plus the cost of your time could well __77__ your hairdryer the most expensive one of all.	68. A) save C) raise B) preserve D) retain 77. A) cause C) leave B) make D) bring

Figure 12: Questions number 68 and 77 of the new CET-4 sample test

4 Conclusion

To assess the usefulness of a large-scale high-stakes test like the CET-4 includes a series of complicated processes that involve varieties of factors and a tremendous amount of work. Fortunately, much research has been done with this regard. The National CET-4 and CET-6 Commission in China has studied the validity and reliability of both the old and new CET-4 tests from a comprehensive perspective and the reports have been published. Other than that, empirical studies, surveys of various kinds can also be found, covering a wide range of issues concerning the different qualities.

This essay studies the six qualities of the CET-4 test based on the test usefulness framework proposed by Bachman and Palmer (1996), by comparing two sample tests which represent the test before and after the reform. However, due to the complicated process, this essay only looks into the test contents and makes a brief investigation into its usefulness. Even so, conclusions and implications can be drawn from the analysis and discussion. Apart from the six qualities, focus is particularly placed on the testing of vocabulary and grammar in the new CET-4 test.

The new CET-4 test is considered to be higher in validity than the old test. To begin with, the new specifications provide more explicit definitions of constructs to be tested as well as how they can be tested. Secondly, the new test covers a wider range of skills to be tested by adopting varied testing techniques in the Listening Comprehension and Reading Comprehension parts. The use of varied test tasks in the listening part, such as Long Conversations and Compound Dictations, along with the original Short Conversations and Listening Passages, engages candidates in different areas of listening ability and corresponds better to the university environment. Similarly, the reading part adopts tests tasks that measure candidates' reading ability in a wider range. Moreover, vocabulary and grammatical structures are not tested in isolation, but are measured within contexts. This is in line with the trend towards pragmatic and integrative testing. Lastly, the way of reporting scores in the new CET-4 test is considered to enhance the validity because a profile of scores not only provides test-takers and test users with the overall language ability, but also indicates the strengths and weaknesses in the abilities that have been measured.

When it comes to test reliability, it is difficult to make a conclusion as to which test is more reliable from the brief comparison between the two sample tests. On the one hand, the old test adopted more multiple-choice formatted questions, which ensured a higher degree of objectivity in scoring, than the new system. On the other hand, the new test, the listening part in particular, introduces more tasks with varied testing techniques, which results in an increased number of items and lengthened time for testing. This change makes the new test more reliable to a certain extent than the old system.

Authenticity and interactiveness measure the correspondence between test tasks and TLU tasks as well as test tasks and test takers, respectively. By an elaborative comparison of Listening, Reading and Writing parts in the sample tests, it is found that both the old and new tests are authentic in a limited degree, but the new test is more authentic than the old one. Firstly, the Listening Comprehension parts in both tests mainly contain functional conversations of daily life whereas few of them take place in the academic setting where students are supposed to use English most often. However, the sample test of the new system

comprises more academic conversations and testing techniques that resemble the academic setting. Secondly, the reading parts in the two tests cover a wide range of topics that are familiar to candidates. Nevertheless, since the new test introduces Skimming and Scanning, two skills often used by students both in the academic setting and daily life, the new test is thus considered more authentic and interactive than the old system. However, we cannot find any passages for practical purposes such as instructions and schedules, which students are most likely to encounter in their future careers. Finally, as there is not much change in testing the writing ability, the authenticity of this part cannot be compared. However, this part bears the same problem as the Reading Comprehension part in that it usually excludes writings for practical purposes.

As a large-scale high-stakes test, the CET-4 inevitably exerts considerable impact on college students, English teachers, the educational system and the society at large. The CET-4 test tends to affect learners' learning styles and contents, and the influence becomes increasingly stronger as the test date draws nearer. The test becomes a motivation for students, initiating students to engage in quantitative and qualitative learning. On the other hand, the high stake it causes puts students under so much pressure that it unavoidably yields negative washback. As for the teachers, the CET-4 test has an influential impact on teachers' curricular planning and instruction. Teachers tend to organize their teaching activities in accordance with the test. Since the new test increases proportions of listening and speaking competences, teachers modify their instructions and curricular accordingly which involve more communicative contents. However, such a change in teachers' instruction and curricular planning is not simply due to the new CET-4 specifications but to the change of teaching materials as well. Finally, the CET-4 test has imposed implicit ideas about success, knowledge, and ability of an individual and thus influences not only the individual, but the society as well.

As a nation-wide test that has been administered for over two decades, the CET-4 is seen as highly practical in that the administrative work is well managed and the educational institutions concerned at all levels work coordinately to ensure the smooth operation of the test, in terms of test development, administration and scoring.

Instead of an individual part to test vocabulary and grammatical structures, the new CET-4 test claims to measure this ability in all other parts, as specified in the specifications. We can easily see traits of grammar and vocabulary testing in the Listening, Reading and Writing parts. Apart from that, two other parts, namely, Banked Cloze and Cloze, also serve this particular purpose, where lexical and grammatical knowledge are tested within a context, which is more compatible with the recent trend towards pragmatic and integrative testing. Nevertheless, Cloze is found to bear several problems when it is used to test vocabulary and grammar: 1) selection of inappropriate passage and deletions might affect the validity and reliability; 2) distracting choices are difficult to write and 3) there might be inequality with distribution of vocabulary items and grammar items.

From the comparison between the two tests and the discussion, it is, therefore, concluded that the new CET-4 test is in a higher degree in terms of the test usefulness than the old test. Also, by comparing the techniques used in the two tests, it is found that the new test is more effective and valid in testing grammar and vocabulary, although it is not without problems.

This essay probes briefly into the usefulness of the new CET-4 test. However, due to the short time since its introduction, further and more elaborative investigations into the new system need to be conducted with regards to the six qualities. What is more, reform with such a large-scale test will bring about response from the society and the people in the system. Therefore, time is still needed for the washback effect to show. Every effort should be made to minimize the negative effect while at the same time maximizing the positive impact. Finally, the reform of the CET-4 test and the syllabus of the College English Course have given rise to fundamental changes in teachers' instructions and curricular planning. Further studies are called for in search of teaching approaches that suit the new requirements.

Reference list

Primary materials

College English Test Band 4 and Band 6 [online] Available from World Wide Web: <www.cet.edu.cn> [Accessed April, 2009]

January 2002 test paper. [online] Available from World Wide Web: <http://cet.hjenglish.com/detail_352.htm> [Accessed April, 2009]

Sample test paper of the new CET-4 test. [online] Available from World Wide Web: <http://learning.sohu.com/20080904/n259379397.shtml> [Accessed April, 2009]

Secondary materials

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

Bachman, Lyle F & Adrian S. Palmer (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

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

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

Carroll, B. J. (1980). *Testing Communicative Performance: An Interim Study*. Oxford: Pergamon Press.

Chen, Feng. (2007). The Washback Effect of College English Test Band 4 on Curricular Planning and Instruction. *CELEA Journal*. Vol. 30, No.1, 2007.

Cumming, A. & D. Mellow. (1995). An Investigation into the Validity of Written Indicators of Second Language Proficiency. [In] Cumming, Alister H. & Richard Berwick (Eds.) *Validation in Language Testing (Vol. 7)*. Bristol: Multilingual Matters Ltd. 72-93.

Ebel, R. L. & D. A. Frisbie (1991). *Essentials of Educational measurement* (5th ed.) New Jersey: Prentice Hall.

Gennaro, Kristen Di. (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. [online]. Available from World Wide Web:

http://journals.tc-library.org/templates/about/editable/pdf/Di%20Gennaro%20Forum.pdf [Accessed May 3, 2009].

Guo, Aiping. (2006a). The Problems and the Reform of College English Test in China. *Sino-US English Teaching.* Vol. 3, No.9, Sep. 2006. 14-17.

Guo, Aiping. (2006b). The Prospect of CET in China. *Journal of Cambridge Studies*. Vol.1, No. 2, June 2006. 39-41.

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

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Miao, Yang (2006). Validating a Simulated Test of CET 4. *Asian EFL Journal*, May 2006, Vol. 12. [online] Available from World Wide Web: <http://www.asian-efl-journal.com/pta_May_06_ym.php> [Accessed May, 2009].

Rea-Dickins, P. (1997). The Testing of Grammar in a Second Language. [In] Clapham, Caroline & David Corson. (Eds.) *Encyclopedia of Language and Education: Language Testing and Assessment* (Vol. 7). Dordrecht: Kluwer Academic Publishers. 87-98.

Read, J. (1993). The Development of a New Measure of L2 Vocabulary Knowledge. *Language Testing*, 10(3):355-371.

Read, J. & C. A. Chapelle. (2001). A Framework for Second Language Vocabulary Assessment. *Language Testing* 18(1): 1-32.

Schmitt, Norbert (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schrutt, N. (1999). The Relationship between TOEFL Vocabulary Items and Meaning, Association, Collocation and Word-class Knowledge. *Language Testing* 16(2): 189-216.

Shao, Hua. (2006). An Empirical Study of Washback from CET-4 on College English Teaching and Learning. *CELEA Journal*. Vol.29, No.1, Feb. 2006. 54-59.

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

Test Reliability and Validity Defined. (n.d.) [online] Available from World Wide Web:http://cc.ysu.edu/~rlhoover/OPTISM/reliability_validity.html [Accessed May 3, 2009].

Wang, Xue-fang, Jian-liang Wang & Feng-fu Liu (2005). CET validity – An Analytical Study Based on the Feedback from the Questionnaires. *Journal of Heibei Institute of*

Architectural Science and Technology (social Science Edition), Vol.22, No.1, March 2005. 109-110.

Wells, Craig S. & James A. Wollack (2003). *An Instructor's Guide to Understanding Test Reliability*. [online] Available from World Wide Web: <http://testing.wisc.edu/Reliability.pdf> [Accessed April 25, 2009].

Wijgh, I. F. (1995). A Communicative Test in Analysis: Strategies in Reading Authentic Texts. [In] Cumming, Alister H. & Richard Berwick.(Eds.) *Validation in Language Testing (Vol.7)*. Bristol: Multilingual Matters Ltd., 154-170.

Yang, Hui-zhong. (2003). The 15 years of the CET and Its Impact on Teaching. *Journal of Foreign Languages*. No.3, May 2003. 21-29.

Yang, Hui-zhong & C. Weir. (1998). Validation Study of the National College English Test. Shanghai: Shanghai Foreign Language Education Press.

Zhan, Ying. (n.d.) *Washback on Chinese learners: An impact study of the College English Test Band 4*. [online] Available from World Wide Web: <http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/164766_Zhan.pdf> [Accessed April, 2009]

Appendix A: Specifications for the CET-4 (Revised Edition) (2006) (Excerpts)

An overview of the CET-4 test

1. Listening Comprehension

This section tests students' ability to obtain information that is presented orally. The listening materials are read with standard British or American English, at a rate of 130 words per minute. The Listening section constitutes 35% of the total score, in which Listening Conversations constitutes 15% and Passages constitutes 20%. The time is 35 minutes.

Listening Conversations includes two parts, i.e. Short Conversations and Long Conversations, both of which are tested in the format of multiple-choice questions. Short Conversations comprises 7-8 dialogues, every of which includes an exchange (where the two speakers speak in turn for once) followed by a question. Long Conversations comprises two dialogues, each of which contains 5-8 exchanges followed by 3-4 questions. Listening Conversations includes 15 questions. Every conversation is read once, with a 15-second interval after the question.

The part of Passages includes Listening Passages (multiple-choice questions) and Compound Dictation. Listening Passages includes 3 passages with 200-250 words in each followed by 3-4 questions, with a 15-second interval between each question. Totally, there are 10 questions. The passages are read once. Compound Dictation tests the listening ability at multi dimensions (from vocabulary to discourse). It includes one passage of 200-250 words, with several words and sentences missing. The passage is read for three times. Candidates are required to fill in the gaps with the information they hear. The words filled in must be the original ones and the sentences filled in can either be the original ones or sentences using one's own words.

2. Reading Comprehension

This section contains two parts, namely, Reading in Depth, and Skimming and Scanning (or Fast Reading). It tests candidates' ability to obtain information in the written form via reading. The section constitutes 35% of the total score, in which Reading in Depth constitutes 25% and Skimming and Scanning constitutes 10%. The time is 40 minutes.

Reading in Depth requires candidates to read three passages, two of which, in the format of multiple-choice questions, contains 300-350 words each. The third passage, Banked Cloze or Short Answer Questions, contains 300-350 words. The part of Reading in Depth tests the reading ability at multi dimensions, including understanding main ideas and important details, comprehensive analysis, making inferences and word-meaning guess with context cues, etc. Each multiple-choice formatted passage is followed by a number of multiple-choice questions, which requires candidates to choose one appropriate response from four choices based on their understanding. Banked Cloze tests vocabulary in a given context. it requires candidates to read a passage which has a number of word deletions and choose appropriate words from given choices so as to restore the original passage. Short Answer Questions contains a passage followed by a number of questions to which candidates should give response or with the shortest possible sentences (less than 10 words) based on their understanding.

Skimming and Scanning comprises 1-2 longer passages or several short passages. the total length is approximately 1000 words. It requires candidates to obtain information with skills of skimming and scanning. Skimming tests candidates' ability to obtain the main idea of a passage at a reading rate of 100 words per minute. Scanning tests candidates' ability to locate specific information by means of various context cues, including numbers, capitalized letters, the first sentence of a paragraph and first word of a sentence, etc. This part adopts formats such as true/false statements, gap-filling sentence completion, and so on.

3. Cloze

Cloze tests candidates' ability of language comprehension and use at multi dimensions. The passage contains 220-250 words on topics that are familiar to candidates. This part constitutes 10% of the total score. The time is 15 minutes.

The cloze passage contains 20 blanks of words which covers either content words or functional words. There should be one word for one blank. There are four choices for each blank, and candidates should choose an appropriate word so that the passage is restored.

4. Writing and Translation

This section tests candidates' ability to produce written discourse in English. It constitutes 20% of the total score, in which Writing accounts for 15% and Translation 5%. The time is 35 minutes.

Writing adopts topics that are familiar to candidates, who should write a passage of no less than 120 words, on a specific topic together with a given situation/outline, graphs/tables, or pictures. Candidates are required to write a passage that can correctly express their ideas, is coherent in meaning, and free of serious grammatical errors. The time is 30 minutes.

Translation contains 5 sentences which candidates should translate from Chinese to English. Every sentence contains 15-30 words. part of the sentence is presented in English, and candidates should translate the missing part (with Chinese meaning presented) according to the meaning of the whole sentence. The time is 5 minutes. The translation should be correct grammatically and pragmatically.

5. Selection of materials

Testing materials are selected from original discourses, including daily-life conversations, lectures, TV or radio programs, newspapers, magazines, books, academic journals, etc. The principles are as follows:

1) Wide range of topics, covering such fields as humanities, social sciences, natural sciences, etc. But the related background knowledge should be familiar to candidates or given in the passages;
2) A variety of styles, including narratives, descriptions, arguments and so on;
3) Medium level for Passage in the Reading in Depth part, and lower level for passages in

parts like Skimming and Scanning, Listening Comprehension and Cloze;

4) Vocabulary within the range specified by College English Course Syllabus (CECS); Chinese meanings or English definitions for important words which are beyond the CECS.

Language skills to be tested and requirements

1. Listening Comprehension

This section tests candidates' ability to obtain information that is presented orally, including understanding of main ideas, important facts and details, implied meanings, making inferences about speakers' opinions and attitudes, etc. The specific skills to be tested are as follows:

A Understanding main ideas and important details

1 identifying central theme

2 listening to important or specific details

3 judging speakers' opinions, attitudes, etc.

B Understanding implied meanings

4 making inferences

5 judging the communicative function of language

C Understanding listening materials with the aid of language features

6 distinguishing phonetic features such as liaisons, stress and intonation

7 identifying relations between sentences, such as contrast and comparison, cause and effect, degrees, purposes, etc.

The Listening Comprehension section of the CET-4 requires candidates to meet the criterion (general level) specified in College English Course Syllabus, namely, able to listen to lectures given in English, daily life conversations, lectures of general topics; able to understand English programs conducted with a medium-speed of 130 words per minute; to obtain the main ideas and important information of such talks; and able to adopt basic listening skills to aid understanding.

2. Reading Comprehension

This section tests candidates' ability to obtain information in the written form via reading, including comprehending main ideas, important facts and details, inferences, judging writers' opinions and attitudes, etc. The specific skills to be tested are as follows:

A Identifying and comprehending main ideas and important details

1 understanding concepts or details that are explicitly stated

2 understanding concepts or details that are implied (such as conclusions, judgments, inferences, etc.); comprehending passage meanings by means of a judgment on communicative functions of sentences (such as requests, refusals, commands, etc.)

3 understanding the central theme of a passage (e.g. locating key summarizing points, and so on)

4 understanding writers' opinions and attitudes

B Comprehending passages using language skills.

5 understanding meanings of words and phrases (such as with context cues)

6 understanding relationship between sentences (such as cause and effect, purposes, comparison and contrast, etc.)

7 understanding connections between different parts (such as by means of lexical and grammatical knowledge)

C Using reading skills such as

8 skimming to obtain main ideas

9 scanning to obtain specific information

The Reading Comprehension section of the CET-4 requires candidates to meet the criterion (general level) specified in College English Course Syllabus, namely, able to read English passages on general topics at a rate of 70 words per minute; able to read longer passages with less difficulty at a rate of 100 words per minute; able to read articles in English newspapers published in China and functional reading materials in a specific future career and daily life; and able to use effective reading skills.

3. Writing and Translation

The Writing section tests candidates' ability to produce written discourse in English. it requires candidates to write English passages to express ideas correctly, with coherent meaning and without serious grammatical errors. The Translation section requires candidates to translate the Chinese in a sentence into English, using vocabulary and structures that are grammatically and pragmatically correct. The specific skills to be tested are as follows:

A Ideas

1 expressing central ideas

2 expressing important or specific information

3 expressing opinions, attitudes, etc.

B Organization

4 writing a narrative, argument or description on the given topic, with prominent key points

5 writing coherent sentences and paragraphs

C Language

6 using appropriate vocabulary

7 using correct grammatical structures

8 using appropriate sentence patterns

9 using correct punctuations

10 using connectives in expressing syntactic relations (such as comparison and contrast, cause and effect, degree, purpose, etc.)

The Writing section of the CET-4 requires candidates to meet the criterion (general level) specified in College English Course Syllabus, namely, able to perform general writing tasks, such as describing personal experiences, feelings and events, giving opinions; able to write passages for practical purposes; able to write a passage of at least 120 words within 30 minutes on a given topic or outline, with intact ideas, proper use of vocabulary and coherent meaning; and able to write with basic writing skills.

The CET-4 test does not test candidates' ability of translation in an individual section. This part primarily tests candidates' ability to express ideas with proper vocabulary and correct grammatical structures.

4. About vocabulary and grammatical structures

In the CET-4, testing of vocabulary and grammar is performed in all sections rather than in an individually-set section. An amount of 4500 words and 750 phrases is necessary in order to meet the requirements specified in each section.

Appendix B: Specifications for the CET-4 (2005) (Excerpts)

The test constitutes five sections: Listening Comprehension, Reading Comprehension, Vocabulary and Structure, Cloze and Writing. All the questions are uniformly numbered.

Part I: Listening Comprehension

There are 20 questions in this part. The time is 20 minutes.

This part contains two sections. Section A includes 10 tasks, each of which is composed of a dialogue, followed by a question. Section B includes 10 questions, which come after a number of listening passages. each passage is followed by 2-4 questions.

There is a 15-second interval after each question. Candidates are required to choose the best response from 4 choices. The dialogues and passages are read once at a rate of 120 words per minute.

Material selection should follow these principles:

1. Dialogues are from daily-life conversations, with less complex sentence structures and topics.
2. Listening passages contain stories, talks, narratives, etc. with familiar topics and les complex scenario.
3. The vocabulary used in this part should be within the range that is specified by College English Course Syllabus.

The Listening Comprehension part aims to test candidates' ability to obtain oral information.

Part II: Reading Comprehension:

There are 20 questions in this part. The time is 35 minutes.

This part contains a number of short passages, with a total of no more than 1000 words. Every

passage is followed by a number of questions. In each question, candidates are required to choose the best response from four choices.

Material selection should follow these principles:

1. A wide range of topics are covered, concerning people, society, culture, common knowledge and basic scientific knowledge, etc.. The background knowledge is familiar to candidates;
2. a wide range of styles are included, such as narratives, descriptions or argumentations;
3. The language used in passages is medium in difficulty. The important words whose meanings cannot be guessed, such as ones beyond the Syllabus, are defined in Chinese.

Reading Comprehension mainly tests the following abilities:

1. to understand main ideas;
2. to understand facts and details related to main ideas;
3. to understand superficial meanings; to make inferences according to contexts;
4. to understand meanings of individual sentences and their relationships with the context.

This part aims to test candidates' ability to obtain information via reading. Correctness and speededness are both required.

Part III: Vocabulary and Structure

There are 30 questions in this part. The time is 20 minutes. 40% of the items tests use of words and phrases, and 60% of the items tests knowledge of grammatical structures. Candidates are required to choose the best response from 4 choices.

Vocabulary and Structure aims to test ability to use words, phrases and grammatical structures. The questions cover vocabulary and grammar specified in the College English Course Syllabus.

Part IV: Cloze

There are 2 questions in this part. The time is 15 minutes. There is a passage of familiar topic and medium difficulty (approximately 200 words), with 20 gaps. One gap comprises a question, with four choices, from which candidates are required to choose the best response based on their understanding of the whole passage, so that the passage becomes intact in meaning and organization. The words filled in the gaps contain content words and function words.

Cloze aims to test candidates' ability to use language comprehensively.

Part V: Writing

There is 1 task in this part. The time is 30 minutes. Candidates are required to write a short passage of 100-120 words. Candidates should write on a given topic, scenario or according to pictures, or with key words, or continue writing according to the first sentences that are given in the beginning of every paragraph. Candidates should express their ideas correctly with coherent meaning and without serious grammatical errors. The topics in this part involve daily life and common knowledge.

Writing aims to test candidates' preliminary ability to express ideas in the written form with English.