*THE JOURNAL OF ASIA TEFL* Vol. 1, No. 1, pp. 75-100, Spring 2004

# *Comparability Study of Two National EFL Tests (CET-6 and TEM-4) in China*

**Yuemei Zhou**
*Shanghai University of Finance and Economics, China*

The two Chinese nation-wide tests (CET-6 and TEM-4) have been in operation for more than a decade during which their status has risen. Nevertheless, there has been little research on the comparability of these two influential tests in China, though both are designed particularly for the second-year students at tertiary institutions. To bridge the gap, this study compares the two tests in terms of test takers, test scores and test content. The results show more overall similarities than differences with a Pearson correlation coefficient of .714 ($p < .01$, nondirectional). Two types of subjects (English majors and non-English majors) in the study were quite similar to each other, with an obvious difference chiefly in instruction hours. The respective means of 67 and 69 appeared close, but this difference was found to be statistically significant. As for test content, differences were observed at a deeper level than at a surface level. The results are discussed in terms how meaningful the relationship is and what they may mean to researchers and test users in developing, using or taking such EFL tests in China.

For about 15 years the Chinese English Test (CET) and the Test for English Majors (TEM) have been in operation nation-wide, and their scores have been increasingly used as measures of proficiency in English as a foreign language (EFL) throughout China. The number of test candidates has increased every year; about 4,500,000 took the CET and about 100,000 the TEM in 2002. The large numbers of test takers along with the extensive use

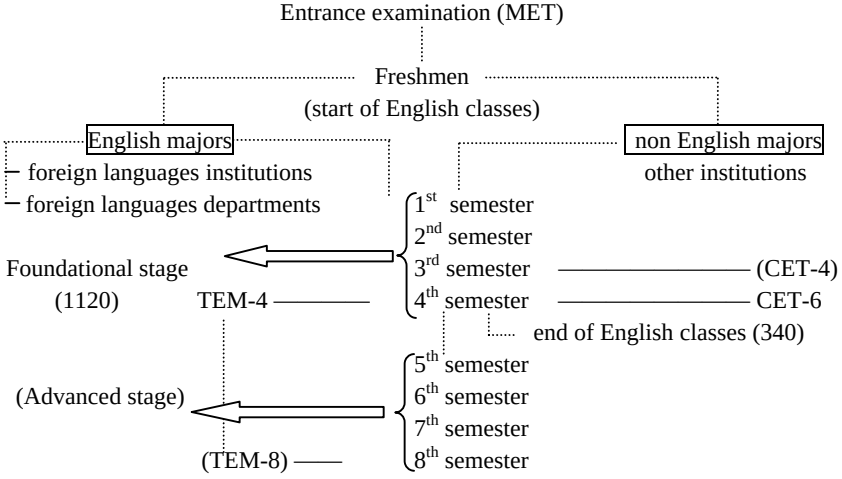
of test scores mean that a great number of careers and education decisions are affected. These decisions include award of diplomas before graduation, applications for admission to an educational program, or seeking employment as well as advancement in a career.

What is the target of the test? The National College English Tests-Band Six (CET-6) was originally designed for college students after they completed the sixth and highest level of study for non-majors (CET designer group, 2000). Some institutions have made the CET-6 certificate a prerequisite for graduation (Huizhong, 2002). The test has been sponsored by the Higher Education Department of the PRC Ministry of Education (Huizhong, 2002) since its inception in 1987. Twice a year the test is administered by the National College English Testing Committee of China (CETC), a testing service centre now based at Shanghai Jiaotong University.

The Test for English Majors-Band 4 (TEM-4) is designed for students majoring in English language and literature; it is given near the end of the first two-year foundation stage of a four-year degree programme. The test is given every May at another testing centre located in Shanghai International Studies University, also under the auspices of the Ministry of Education.

As each test is claimed not to be solely limited to certain textbooks, they are in essence proficiency tests (Huizhong, 1998; Shen, 1998-b), though devised in accordance with the requirements of the respective national EFL teaching syllabuses (Huizhong, 2002; Shen, 1998). The only distinction between the two syllabuses seems to lie in the fact that the English majors have much more class instruction (about 1120 class hours) than the non-English majors (about 340 or even less). The fourth semester is the normal time for the majority of both types of students to take their respective test. CET-6 test takers need to take a pre-intermediate form of the test (CET-4); TEM students can take an advanced form of the test (TEM-8) two years later. Figure 1 illustrates this.

**FIGURE 1**
**Relationship between School Years and English Tests**




The entrance examination, known also as MET refers to the yearly national test of English, among the other summative tests of Chinese language, chemistry or mathematics for high school graduates, planning to enter colleges and universities. Only those who achieve acceptable scores in these examinations are admitted to major in English or in other humanities and science disciplines. Four semesters later, they are expected to take the TEM-4 or CET-6 tests.

In recent years, however, the CET-6 has been taken also by English majors, while the TEM-4 also attracts non-English majors in some parts of China. Thus the distinction between the two tests has become blurred, and they seem to be practically regarded as something equivalent or at least interchangeable. Until now, no official or research explanation for the exchange has been presented. It was this phenomenon that aroused the author's curiosity to explore the relationship between the two tests. This interest has now resulted in a research undertaking of the present study.

The present study, therefore, attempted to answer the following questions. First, is there anything in the tests themselves to make these two tests

interchangeable? Are they related, or are they different? Further, what can be done to further manifest or blur the distinctiveness of either test? As a preliminary study, this paper intends to cover only the first two questions after a brief summary of the associated theories and research literature on language ability and previous comparability studies.

## LITERATURE REVIEW

As a theoretical as well as empirical background, this literature review gives a brief review of the communicative language ability (CLA) and test method facets (TMF) theories that are commonly considered to be fundamental in test assessment. The ideas or concepts from this theoretical review will be used to describe or analyze the test content in the subsequent comparison of the two tests. Then characteristics of some research works concerning comparability study are discussed, with a view to establishing the operational framework called comparability-oriented model (COM) (Yuemei, 2003) for the present study.

### Communicative Language Ability (CLA)

The CLA model, according to Bachman, is composed of five elements. Language competence (LC), strategic competence (SC) and psychophysiological mechanisms (PM) are three elements that constitute the core of the CLA, while the other two elements of world knowledge and context of situation function to render language competence dynamically communicative (Bachman, 1990, p. 85). To facilitate the description of the study, the focus of review is chiefly on LC and PM (Bachman, 1990):

**LC** Language competence is usually viewed at two levels: the level of organizational competence and the level of pragmatic competence.

**PM** The ingredients involved in the PM are essentially the neurological and physiological processes occurring when we human beings execute language.

Details about LC and PM are briefly listed in Table 1.

**TABLE 1**
**Components of Communicative Language Ability (CLA)**

| | | | |
|---|---|---|---|
| Knowledge of the world | | | |
| Language competence (LC) | 1. Organizational competence | a. grammatical | 1) vocabulary<br>2) morphology<br>3) syntax<br>4) phonology/graphology |
| | | b. textual | 1) cohesion<br>2) rhetoric |
| | 2. Pragmatic competence | a. illocutionary | 1) heuristic<br>2) imaginative<br>3) ideational<br>4) manipulative |
| | | c. sociolinguistic | 1) dialect<br>2) register<br>3) naturalness<br>4) culture/figures of speech |
| Strategic competence (SC) | | | |
| Psycho-physiological mechanisms (PM) | | 1. Receptive mode | a. listening (aural)<br>b. reading (visual) |
| | | 2. Productive mode | a. speaking (oral)<br>b. writing (visual) |
| Context of situation | | | |

Later, for language testing in particular, this general CLA model was slightly revised by Bachman and Palmer (1997, p. 63), in which the role of test takers is seen to be even highlighted, though termed otherwise. Dotted lines in Figure 2 show this between CLA in general communication and CLA in testing.

**FIGURE 2**
**Comparison of Two Models of CLA in Use and CLA in Testing**

| CLA in general use | | CLA in testing |
|---|---|---|
| Knowledge of the world | | Topical knowledge |
| Knowledge of language | | Language knowledge |
| Strategic competence (SC) | - - - - - - - - - - - - - - - | Strategic competence |
| Psycho-physiological mechanisms (PM) | - - - - - - | Personal characteristics |
| Context of situation | | Test tasks and test setting |

From Figure 2, the central position of human beings or test takers in successful language communication in general as well as in testing is theoretically brought out. In Bachman's words, the first four components in either the left or right column in Figure 2 "represent characteristics of individual language users," while the last component "includes characteristics of the task or setting with which the test takers interacts" (Bachman & Palmer, 1997, p. 62).

Of the preceding four, the component of personal characteristics usually refers to "individual attributes that are not part of test takers' language ability but which may still influence their performance on language tests" (Bachman & Palmer, 1997, p. 64). These individual attributes may vary from the most common (such as age or gender) to less obvious ones (such as cognitive style). According to Cohen (1994), personal characteristics in the context of language testing include age, foreign language aptitude, socio-psychological factors, personality, cognitive style, language use strategies, ethnolinguistic factors, and multilingual ability. While Bachman and Palmer's (1997) list of personal characteristics includes 1) age, 2) gender, 3) nationality, 4) resident status, 5) native language, 6) level and type of general education and 7) type and amount of preparation or prior experience with a given test.

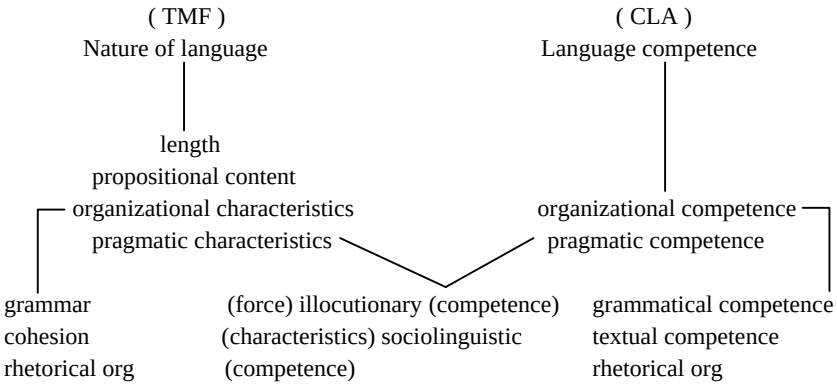
This study chiefly follows Bachman, with some adaptation to local conditions though, for these attributes are thought to be largely predictable and better understood (Bachman & Palmer, 1997) than a great number of other personal characteristics that could potentially affect the test performance of any given test taker.

## Test Method Facets (TMF)

Similar modifications have also been made to the TMF model in the study. According to Bachman (1990, p. 119), TMF includes five major categories: 1) the testing environment, 2) the test rubric, 3) the nature of the input the test takers receives, 4) the nature of the expected response to that input and 5) the relationship between input and response. Each of these categories has its set

of sub-categories (see Bachman, 1990 for details). Figure 3 gives a sketch of TMF and CLA categories combined.

**FIGURE 3**
**Echoes of Language Facets in TMF and CLA**




To sum up, both CLA and TMF start from concerns that affect test performance and each of them ends up with a list of characteristics. They can be seen as one theory of testing with two components (Bachman, 1990), which provide the applied linguistic foundation that informs the following discussions.

## Previous Comparability Research

In addition to the CLA and TMF models, the present study also derives operational support from empirical evidence of some comparability research, on the basis of which an operational model has been proposed.

Generally speaking, research on test comparability may differ, at home and abroad, in form as well as in content. Some researchers find interest in comparing tests of listening, reading or writing; others find more interest in vocabulary (Beglar & Hunt, 1996), cloze, or grammatical parts in two comparable tests; still others focus on comparisons of scoring methods,

performances in general or field-specific tests or misfitting items.

**TABLE 2**
**Procedures Taken by the Comparability Studies on Standard Tests**

| Step/components | Between CET-4 and CET-6 (reported scores) | Between CET-4 and CET-6 (weighted scores) | Between TEM-4 and TEM-8 (weighted scores) | Between TOEFL and FCE (standard scores) |
|---|---|---|---|---|
| A. subjects | The same sample group of 426 from 7 universities to take CET-4 in the morning and CET-6 in the afternoon of the same day. | | The same sample group of 114 to take TEM-4 first and to take TEM-8 two years later. | 1) pilot study: a two-level sample group of 259 to take the tests during a three-week period at 3 sites.<br>2) a sample about 1400 to take the two tests on adjacent days at 8 sites. |
| B. scores | to compare **means** and standard deviations (SD) | also to check **mean** difference for each component with F Test | to compare the **characteristics of** the two-set **scores** | to compare **means**, standard deviations with the norms |
| | to observe score **distribution** | | | |
| | to examine **correlation** coefficient | also to observe **correlation** matrix | | to observe **correlation** matrix |
| | | | | to estimate test **reliability** |
| | | | | to investigate the **factor** structure |
| | | | | to determine **score interchangeability** |
| C. content | | | | to analyze the content with ratings |
| D. others | | | | to examine test familiarity or preparation |
| E. report | Conclusion or discovery | | Conclusion or discovery | Implication and recommendation |

Of these cases, two types of comparability study on testing are especially relevant. The first type is concerned with large-scale standardized EFL tests, while the second type deals with specific aspects of language competence or school based tests. The following two tables offer brief outlines of the research procedures used by six research studies (Beglar & Hunt, 1996; Huizhong & Weir, 1998; Schmitt, Schmitt & Clapham, 2001; Yuemei & Yanli, 2002) in China and abroad, in order to provide some operational background for the present study.

These studies are all characteristic of comparison between two or more than two tests. The first type of comparability study (in Table 2) reveals that all the three cases - with information respectively derived from the published reports (Huizhong, 1998; Shen, 1998) – follow a certain process or pattern of research. As the three studies are generally regarded as successful validation projects, the patterns in each are likely to be useful to other researchers undertaking a comparability study.

The top row shows that all the three cases compare two tests taken by one sample group. Though the first two are not as closely associated with test comparison as the third one, all three take certain steps in conducting the research project. Looking at the steps in the left column, five can be clearly identified as forming the pattern of comparability research. Step A is an initial step concerning sample subjects of the group. As following steps, Step B is score analysis, and Step E reports what has been discovered. These three appear to be the most frequently employed procedures in exploring test comparability. Full-length research on test comparability (like the study comparing TOEFL and FCE) may involve two additional steps. Step C dealing with test content (chiefly qualitative) and Step D concerning related issues like test preparation. As Step B contains a number of statistical techniques, the key words are bold-faced not only to facilitate comparison across the columns in the table but also to be of convenience for later discussions. Now let us look at the second type of comparability study in Table 3.

**TABLE 3**

**Procedures Taken by the Comparability Studies on Specific and School Tests**

| Step/component | Between 4 forms within each of 2 different vocabulary tests | | Between 2 college-wide achievement tests | Between 2 new versions of vocabulary tests |
|---|---|---|---|---|
| | Original | Revised | | |
| A. subjects | The one sample group of 496 to do one test in 4 forms; the second sample group of 464 to take the other test in 4 forms, in order to create two new forms for each test. | | The one sample group of 1185 doing one test in the morning; the second sample group of 976 doing the other test in the afternoon of the same day. | The two tests combined into one to be taken by a sample group of 801 in 13 groups at 9 sites in 5 countries. |
| B. scores | to compare **means**, SD, range, and SEM | to compare **means** difference using t-test | to compare **mean**, variance, and covariance to show the equivalence of the two tests | to compare **means**, SD, skewness, kurtosis, minimum, maximum, range etc. also with the norm |
| | to observe **score distributions** | to compare score **distribution** variances using F-test | | to compare **score distributions** |
| | to compare **reliability** estimates | to compare **correlation** coefficients | to compare **reliability** estimates | to compare **correlation** coefficients, also with the norm |
| | | to examine eigenvalues and **factor** structures | to compare **factor** analysis | |
| | to compare good items using **item**-total correlations | | to conduct **item** analysis (difficulty level and discrimination index) | to compare **item** indices (difficulty level and discrimination index) also with the norm |
| | | | **Component** analysis or section profile | |
| C. content | | | | |
| D. others | to compare these coefficients with TOEFL | | | 1. pilot study<br>2. native speaker results |
| E. report | Two newly-revised test forms | | Comments on the two tests | Conclusion |

The first case appears to be a multi-faceted and detailed investigation of two vocabulary tests, as it sets to "carefully analyze and validate the revised versions" (Beglar & Hunt, 1996, p. 131) developed for high school and university students in Japan. The second "reports on a study which uses a range of analysis techniques to present validity evidence, and to explore the equivalence of two revised and expanded versions of the original tests" (Schmitt et al, 2001, p. 55) designed for international learners of English for general or academic purposes. The third concerns the quality of two school-based tests for non-English majors in China (Yuemei & Yanli, 2002). The focus of attention in these cases is somewhat different when compared with the studies mentioned earlier, but their research procedures appear similar, ommiting analyses of test content.

There are at least two points that may draw our attention to these procedures:

1. analyses of content are made in large-scale national or international EFL tests but absent in local ones;
2. subjects, scores and test content seem to be common components of these comparability studies

On the basis of these observations, the author proposes a model for carrying out comparability research in testing. In order to distinguish comparability study from other validation research on language tests, this model is called *comparability-oriented model* (COM), meaning no matter how far or profound the study goes, the attention is chiefly directed to the comparative aspects. It consists of three major components: sample subjects, scores, and test content. While a project might omit a component, studies with all three present tend to be more persuasive.

The author does not presume to present this model as a perfect framework. The COM serves as a guide or pointer, as it were, to chart directions for comparability research. As research advances, it is likely that improvement will be made in the framework. Instead of something rigid, therefore, this

COM model is meant to be flexible, adaptable and heuristic, a starting point for comparative inquiry into language tests that will constantly reshape itself, to dynamically enrich its analytical power.

## THE STUDY

### Methods and Context

The following discussion will be carried out through analyses of the three components of comparability study, that is, test takers (or sample subjects), test scores, and finally test content, very briefly in this paper, backed up by the operational framework of COM. Data used in the study were of three types. The first type of information was data about the test takers, obtained from a questionnaire when they finished the CET-6 paper. The second type of data was the test scores on the two tests. The third type of data came from the teachers' ratings of test content. These data were examined and analyzed, quantitatively as well as qualitatively where necessary. All the multi-facet analyses aimed at exploring the relationship between the two tests, if any, in a relatively objective and comprehensive way. Finally, the findings were summarized, with reference to CLA and TMF theories.

Along with the process of analyses, some important assumptions were made, including:

1. The sample subjects are homogenous by nature for each test, and the scores on each test are thus comparable.
2. The norms to be used in mean comparison are reliable, as they are officially published data, with the CET norms from the 1995 administration and TEM-4 norms from the 2002 testing population.
3. The normal distributions and equal variances of the scores are satisfied (Brown, 1995, p. 166).
4. No significant relationship exists between the two tests, hence the null

hypothesis.

There were also some considerations in selecting the sample test papers for the study. The CET-6 paper was the one used in June 1996. One reason for this selection is that students usually pay more attention to recent test papers than those used long ago, and even if some of them had seen or done it before, they might not have a fresh memory of its content. Another reason is that no easy contact for using the currently administered test was available to us.

With regard to the first reason, the study sample were then asked about the degree of their previous exposure to the CET test paper in the questionnaires. The result showed (Table 4) that a similar percentage of each group (54%; 50%) had never seen or done the sample test before. This not only confirmed, to some degree, the expected closeness of the two sample groups of the study, but also helped explain our control over the variable of test familiarity.

For the other test, we were luckily able to administer the current paper of the TEM-4 in the year of 2002, thanks to the help and permission given by the TEM testing centre.

Considering the initiative of the study, any versions of the two tests could serve our purpose of comparing them, as each of the test papers should have gone through professional procedures of design, development and investigation of reliability and validity by the testing authorities. One sample of each can thus represent the relevant test to a large extent. The two tests were then administered to the combined sample subjects, together with the questionnaires, one after the other during the course of a few weeks at two sites far apart from each other.

## Subjects

Test takers have various personal characteristics which “may … influence their performance on language tests”, according to Bachman and Palmer (1997, p. 64). Information about the sample test takers here is to show characteristics and the degree to which they represent the larger population of

Chinese students.

The sample subjects were chosen from two universities of different types in the metropolitan city of Shanghai. One is a leading institution in foreign language studies and the other is a university of finance and economics. Both enroll excellent students, whose college majors are different. These students may represent two types of students as well as test takers two years later. Four classes of currently registered sophomores were chosen from each on a random basis, for efficiency and convenience as well (Hanliang, 2000). The reasons for this sample selection from the universities reflect these considerations:

1. The sample is as much characteristic of the operational populations of the CET and TEM as possible in terms of majors or years of English study at college. The subjects need to be a mixture of two different types of institutions or departments.
2. Meanwhile, the two groups of selected subjects should have similar background, such as age or prior learning of English, so as to build up a reasonable basis for meaningful contrast and comparison of their characteristics (similarities).
3. It is possible for the sample subjects to take the two tests one after the other within a short time and under conditions as similar to normal administrations as possible.
4. Each group is expected to be familiar with the respective type of the tests that is particularly designed for them. That is, if one candidate takes both tests, he is presumably familiar with one test and unfamiliar with the other.

The following data (in Table 4) give a general picture about the sample subjects for the study. The information obtained from the questionnaires included academic status, the size of each sample group, age, gender, scores from the entrance examination, attitudes towards English classes and motivation for English learning.

**TABLE 4**

**Test Takers' Characteristics**

**Size of each sample group (1)**

| | Language status | N. for CET | N. for TEM | N. of sample used for analysis |
|---|---|---|---|---|
| Group A | Eng. majors | 105 | 105 | 85 |
| Group B | non Eng. majors | 130 | 133 | 85 |
| Total | | 235 | 238 | 170 |

**Current age (2)**

| | Eng-maj. | non Eng-maj. | average |
|---|---|---|---|
| Mean | 20 | 20 | 20.014 |
| Median | 20 | 20 | 20 |
| Min | 17 | 19 | 17 |
| Max | 21 | 21 | 21 |

**Gender(3)**

| | Eng-maj. | non Eng-maj. | average |
|---|---|---|---|
| Male | 22% | 44% | 33% |
| female | 78% | 56% | 67% |

**Scores from the Entrance Examination (4)**

| | |
|---|---|
| Group A | 86.31 |
| Group B | 77.17 |
| Average | 81.74 |

**Attitudes to and motivation for English learning (5)**

| ***1. Do you like English?*** | Yes (%) | No or other (%) |
|---|---|---|
| Group A | 76 | 24 |
| Group B | 74 | 26 |

***2. What is the purpose for you to learn English well?***

| | hobby | Knowledge | future job | graduation | passing tests |
|---|---|---|---|---|---|
| | | | % | | |
| Group A | 21 | 49 | 22 | 4 | 5 |
| Group B | 10 | 50 | 25 | 9 | 7 |

**CET familiarity or preparation (6)**

| | Eng-maj(%) | non Eng-maj(%) | All(N=219) |
|---|---|---|---|
| None | 54.3 | 50 | 52.1 |
| Listening once done | 1.9 | 3.5 | 2.7 |
| Reading once done | 3.8 | 14.9 | 9.6 |
| Writing once done | 2.9 | 5.3 | 4.1 |
| More than one | 1 | 1.8 | 1.4 |
| All once done | 36.2 | 24.6 | 30.1 |

The size of each sample group started at 238, but was later reduced to 170 for reasons of incomplete information. The median age of the sample students was 20; two thirds of the sample are female (67%). Their average score on the entrance examination was close to 82. More than two thirds (76%, 74%) of them said they love to learn the English language, and nearly half (49%; 50%) of them thought learning English well would bring them good jobs in the future.

The hypothesized representativeness of the sample groups were statistically examined by the independent or paired *t* tests of score means. That is, the three score means on the CET-6 of the entire sample, Group A and Group B were first compared with another three CET-6 mean norms (or norm means) of 1995 (Huizhong, 1998), and then the three score means on the TEM-4 were compared with three other TEM-4 mean norms of 2002 (offered by the TEM centre). These norms were obtained from officially published reports of test validation. The results appear in Table 5.

**TABLE 5**
**Sample Representativeness of the Operational Populations**

| | CET | TEM |
|---|---|---|
| Entire sample | representative | representative |
| Group A | less representative | less representative |
| Group B | representative | representative |

Considering the statistic significance in mean differences, our sample as a whole appeared to be typical of the test population, both for the CET-6 paper or for the TEM-4 test. Separated, Group A was less representative, which might be attributed to the unusually high proficiency of this group of English majors.

## Score Analyses and Findings

It was more reasonable to group the scores by test paper than by student type, the author believed, as the chief concern of the study was to see

relationship between the two tests rather than distinction between the two types of test takers. So the two sets of scores were comparable.

Based on the data, the multi-facet analyses of scores yielded meticulous but interesting findings. For the sake of simplicity, at least 4 points can be brought forward as major findings.

First, there was a 99% probability (**) that the two sample tests were significantly correlated with a coefficient of .712, thus excluding the null hypothesis of no relationship between the two tests. Another interesting aspect of their relationship, can be seen from a third comparison of means with the MET test (Table 6). That is, they were both significantly correlated with the English examination taken by middle school students before entering college (MET), with a slightly higher coefficient with the TEM-4 (.66) than with the CET-6 (.527), though.

**TABLE 6**
**Correlation between CET, TEM, and MET**

| | CET | TEM | MET |
|---|---|---|---|
| CET | 1.000 | | |
| TEM | .712** | 1.000 | |
| MET | .527** | .661** | 1.000 |

Second, the average means of the tests looked quite close with only a 2-score difference in between (CET: 67.10; TEM: 69.22), indicating a score similarity (Table 7). Statistically, however, the two means were significantly different from each other, with the CET-6 mean a little lower and therefore a little more difficult. Strangely, however, the values showing difficulty level of either test appeared to be confusing, by which the TEM-4 (.698) seemed to be more difficult than the CET-6 (.707). This was confirmed by the students in the questionnaire: about 23% of them thought the CET-6 was more difficult while 35% found the TEM-4 more difficult.

**TABLE 7**
**Statistics about the Scores**

| | | Mean | Std dev | Median | Min | Max | ANOVA F | ANOVA Sig. | Level of difficulty |
|---|---|---|---|---|---|---|---|---|---|
| CET weighted score total | 100 | 67.10 | 12.48 | 68.50 | 20.00 | 88.50 | 71.161 | 0.000 | .707 |
| TEM weighted score total | 100 | 69.22 | 15.94 | 70.00 | 13.00 | 92.00 | 450.233 | 0.000 | .698 |

Third, taking only multiple-choice items of both tests into consideration, we find (in Table 8) from factor analyses that about 27% of CET multiple-choice items and about 77% of TEM multiple-choice items met the criterion for unidimensionality (Reckase, 1979), showing a striking difference in latent structures of discrete items between the two tests.

**TABLE 8**
**Results from Factor Analysis about Dimensionality**

| | **%* of the part** | **% of var**** | | **% of the part** | **% of var** |
|---|---|---|---|---|---|
| CET-6 | | | TEM-4 | | |
| Listing items | 20% | 20.776 | Listing items | 15% | 21.868 |
| Reading items | 40% | 12.930 | Careful reading | 15% | 22.669 |
| Vocabulary and grammar | 15% | 17.472 | Fast reading | 10% | 39.912 |
| | | | Vocabulary and grammar | 15% | 18.067 |
| | | | Cloze | 10% | 20.140 |
| Total | 75% | | | 65% | |
| % of obj. item total | 0.267 | | | 0.769 | |

* percentage of the component in the entire test paper
** Percentage of variance

Lastly, percentile observations (Table 9) reveal that excellent students would probably achieve higher scores on TEM-4 than on CET-6, while students below the average might get lower scores on TEM-4 than on CET-6, with the TEM-4 test thus appearing a bit more discriminative. (This could also be reflected by the values of standard deviations).

**TABLE 9**
**Percentiles of the Sample Score Totals**

| %1 | 5 | 10 | 25 | 50 | 75 | 90 | 95 | average |
|---|---|---|---|---|---|---|---|---|
| C-tot | 40.78 | 47.50 | 61.00 | 68.50 | 76.50 | 81.50 | 84.18 | 65.7 |
| T-tot | 42.55 | 46.00 | 56.75 | 70.00 | 83.00 | 87.95 | 89.00 | 67.9 |
| N=170 | | | | | | | | |
| T-C= | 1.77 | -1.5 | -4.25 | 1.5 | 6.5 | 6.45 | 4.82 | 2.2 |

In brief, of these 4 aspects, the first two show surface similarity and last two show essential distinction. If we just look at the scores, the two tests appeared close to each other, but in essence they are largely different.

## Content Analyses and Findings

Seven EFL teachers were asked, after being trained, to judge the test content with a rating table that was partly derived from Bachman's theory of CLA and TMF. The results of the rating went through a complicated process of analysis. Combined with other analytical devices, findings deal with two aspects: quantity and quality.

### *Quantity*

Let's first have a brief look at the formation or construction of the two tests (Table 10), along with the quantity of input involved in each test, which needs to be processed within the given time: the CET-6 is 120 minutes and the TEM-4 is 140 minutes.

Table 10 sums up the quantitative information for the two tests, including the word count of every part in either test, the input quantity for each test, as well as the weighting of parts and number of items. On the whole, the CET-6 test offered a little more input in quantity (5308 > 5220) than the TEM-4 for test takers to process in the limited time, thus yielding a somewhat greater average of input process per minute (44.2>37.3). This was greatly contributed by the darkened three out of the five parts in the CET-6. In this

sense, the CET-6 appeared to test a higher level of language proficiency.

**TABLE 10**
**Overall View of Test Input Quantity of the Two Tests**

| CET components (120 minutes) | % | # of item | # of Sub-total | TEM components (140 minutes) | % | # of item | # of Sub-total |
|---|---|---|---|---|---|---|---|
| Listening | 20 | 20 | 1730 | Listening | 25 | 25 | 1463 |
| Reading | 40 | 20 | 2462 | Reading (+fast reading) | 15 | 25 | 2808 |
| Vocabulary & Structure | 15 | 30 | 632 | Vocabulary & Grammar | 15 | 25 | 280 |
| Short-answer question | 10 | 10 | 364 | Cloze | 10 | 15 | 305 |
| Composition | 15 | 1 | 120 | Composition | 15 | 1 | 150 |
| | | | | Note-writing | 5 | 1 | 60 |
| | | | | Dictation | 15 | 1 | 154 |
| Total quantity | 100 | 81 | 5308 | | 100 | 93 | 5220 |
| Average (per minute) | | | 44.2 | | | | 37.3 |

On the other hand, we see the TEM test offered four out of the seven components that contained a greater quantity of input. Since there were more components in the TEM test, its tasks appeared to achieve a greater variety, which might make the test takers feel it more challenging.

*Quality*

There are at least three major findings that can be drawn from percentage and content ratings concerning the test papers. The ratings might reveal some professional comments on the test content.

First, of ten general aspects or indicators (Yuemei, 2003) of input characteristics in terms of CLA and TMF, four of them on the CET and six on the TEM showed greater complexity (Table 11). This observation might provide a qualitative approach to understanding the characteristics of input concerning its construction, quantity and nature.

**TABLE 11**
**Content Characteristics of the CET and TEM Tests**

| Content characteristics | Per item. | CET | TEM | Greater complexity (>) | |
|---|---|---|---|---|---|
| I Construction | min per obj. item | 1.057 | .944 | T>C | |
| | % of obj. item | 75 | 65 | T>C | |
| II Length or quantity | words treated per min. | 44 | 37 | | C>T |
| III Nature of language | % of difficult tasks (+) | 30 | 34 | T>C | |
| ---task type | % of easier tasks (-) | 65 | 56 | T>C | |
| ---sentence type | % of simple sentence | 40 | 44 | | C>T |
| | % of complex sentence | 51 | 46 | | C>T |
| ---information type | % of humanities topic | 63 | 90 | | C>T |
| | % of exposition | 38 | 70 | T>C | |
| | % of narration | 25 | 0 | T>C | |

Second, CET-6 did keep a balance in measuring both organizational (2.42 on a scale of 0 to 4) and pragmatic (2.41) language abilities, according to the theory of CLA, while TEM-4 concentrated more on organizational (2.61) abilities than on pragmatic (2.51) language abilities (Table 12). Relatively, the overall level indicated by the figures for measurement power of language ability was a bit higher in the TEM-4 than in the CET-6.

**TABLE 12**
**Average Ratings in Terms of CLA for the CET and TEM Tests**

| Communicative Language Competence | | Average | | | |
|---|---|---|---|---|---|
| | | CET | | TEM | |
| Lexicon & syntax | Organizational | 2.60 | 2.42 | 2.77 | 2.61 |
| Cohesion & rhetoric | | 2.24 | | 2.46 | |
| Heuristic & imaginative | Pragmatic | 2.27 | 2.41 | 2.31 | 2.51 |
| Dialect & register | | 2.56 | | 2.70 | |
| Strategic competence | | 2.00 | | 2.23 | |

Lastly, the differences between the two tests (though fewer than the similarities) might be theoretically traced from CLA and TMF, two major factors or perspectives that have important impact on test performance

(Bachman, 1990). Given the two factors, five CLA indicators and seven TMF aspects were rated for each part of the tests by the teachers. The results show (Table 13) that the two tests were only 28% different in their function of testing language ability, and 53% different in terms of test method facets.

**TABLE 13**
**Summary of Paired Rating Comparisons for both CLA and TMF**

| | CLA | TMF |
|---|---|---|
| Listening | 1/5 | 4/7 |
| Reading | 0/5 | 6/7 |
| Vocabulary and structure/grammar | 2/5 | 0/7 |
| Writing | 1/5 | 2/4 |
| Short answer question + cloze | 3/5 | 5/7 |
| Total | 7/25 | 17/32 |
| % of difference | 28 | 53 |

## DISCUSSION AND CONCLUSION

Analyses could be conducted of many more aspects of the two tests, and the above discussions deal only with some. Even so, many interesting facts were found. Table 14 (C for the CET-6 tests and T for the TEM-4) summarizes the exploration, with similarities termed as closeness and differences as distinction. Virtually no absolute differences were observed between the two tests; every aspect of the one compared was somewhat related to the other, but there was a variation in degree.

**TABLE 14**
**Summary**

| | Surface closeness (more or less) | Latent distinction |
|---|---|---|
| **Scores** | 1. (of the two tests) highly correlated<br>2. significantly correlated with MET<br>3. T- 2-point higher mean<br>4. T- higher difficulty value<br>5. T- more discriminative | C- 27% unidimensional<br>T- 77% unidimensional |

| | Quantity closeness | Quality closeness | |
|---|---|---|---|
| **Content** | C-more input<br>T-greater variety | T-greater | complexity from input characteristics |
| | | T-more | powerful in testing organizational and pragmatic competencies |

| Expert judgement | Similar | Different |
|---|---|---|
| CLA | 72% | 28% |
| TMF | 47% | 53% |

In other words, similarity was generally obvious while difference was seldom outstanding. Concerning scores, on the one hand, a striking distinction was only found in the discussion of latent dimensionality from factor analyses, with the other aspects closely similar at the surface level. On the other hand, nearly no sharp distinction between the two tests was noted in test content.

This was finally confirmed by the expert judgement. Nearly three fourths (72%) might account for the similarity of the two tests in terms of CLA. That is to say, if we wanted to measure language proficiency, either test could serve the purpose, regardless of the one fourth (28%) of difference. In terms of test methods, the percentage of difference between the two tests increased to 53%. All this might manifest the fact that the two tests had a similar power to measure English proficiency, with the TEM-4 perhaps a little more specialized in testing language ability.

## REFLECTIONS AND IMPLICATIONS

Under such a three-perspective framework or model of test takers, test scores and test content, we have found that there was indeed some justification for the tendency that the test candidates take the two tests interchangeably (as mentioned in the Introduction). We are fully aware, of course, that findings are still limited and based solely on the data that has been obtained from a small number of students, though sufficiently representative. Research needs to be continued as to provide more evidence of support or argument.

For the time being, however, these findings did justify, to a larger extent, the interchangeability in using the two tests. Then, one may ask, why are there two forms of similar national tests in China? To answer this question is certainly beyond the scope of the present study. As Spolsky put it in 1995, moreover, any existing phenomenon in the field of language testing may have its historical, sociological and political context. What the author wanted to find out about these tests was the actual relationship between the two tests. With all those findings now, the three research questions (see Introduction) now have their answers. Yes, these two tests are rationally related, both in their power to measure the language ability and in terms of test methods. Under the general similarity between the two, meanwhile, some difference was also revealed, indicating that the two tests were not identical.

As for implications drawn from these findings for future testing practice, efforts could be made to either merge the two tests into a new unity or make them manifestly different. The first suggestion may help practice economy in testing practice, while the second may bring back the confidence of the English majors. After all, they have spent many more hours (1120) in class than non-English majors (340), and a similar score on the TEM-4 to that on the CET-6 could be hardly welcomed.

## THE AUTHOR

Yuemei Zhou is an associate professor of English, working in the Foreign Languages Department of Shanghai University of Finance and Economics in China, with language testing as her research area. She obtained her doctoral degree from Shanghai International Studies University in 2003.

## REFERENCES

Bachman, L., & Palmer A. (1997). *Language testing in practice.* Shanghai: Shanghai Foreign Language Education Press.

Bachman, L. (1990). *Fundamental considerations in language testing*. Shanghai: Shanghai Foreign Language Education Press.

Beglar, D., & Hunt, A. (1996). Revising and validating the 2000 word level and university word level vocabulry tests. *Language Testing, 16*(2), 131-162.

Brown, J. (1995). *Understanding research in second language learning.* Cambridge: Cambridge University Press.

CET designer group. (2000). *Syllabus and sample test for college English test – Band 6.* Shanghai: Shanghai Foreign Language Education Press.

Cohen, A. (1994). *Assessing language ability in the classroom*. New York: Heinle and Heinle.

Hanliang, L. (2000). *An introductory course to statistics.* Shanghai: Shanghai Finance and Economics University Press.

Huizhong, Y. (2002). *The 15 years of the CET and its impact on teaching.* Paper presented at the 1st International Conference on Language Testing and Language Teaching, Shanghai, China.

Huizhong, Y., & Weir, C. (1998). *CET validation study*. Shanghai: Shanghai Foreign Language Education Press.

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Measurement, 4*, 207-230.

Schmitt, N., Schmitt, D., & Clapham C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing, 18*(1), 55-88.

Shen, Z. (1998a). *TEM validation study*. Shanghai: Shanghai Foreign Language Education Press.

Shen, Z. (Ed.). (1998b). *English language testing – Some theoretical and practical considerations.* Shanghai: Shanghai Foreign Language Education Press.

Spolsky, B. (1995). *Measured words.* Oxford and Shanghai: Oxford University Press and Shanghai Foreign Language Education Press.

Yuemei, Z. (2003, June). *Comparability study of two national EFL tests in China*. Unpublished doctoral dissertation. Shanghai International Studies University, Shanghai.

Yuemei Z., & Yanli, Z. (2002, June). A case study on the quality of school-based CET. *Foreign Language World*, 71-78. Shanghai: Shanghai Foreign Language Education Press.