Language Testing 2008 25 (1) 15–37

The key to success: English language testing in China

Liying Cheng *Queen's University, Canada*

The testing and examination history in China can be traced back to the imperial period nearly two thousand years ago. The existence of English language testing (tests), on the other hand, has a much shorter history. These English tests, developed and administered over the past 20 years, however, are taken by billions of learners of the English language in China. To many of these learners, doing well on these tests are the key to their academic success as well as the success of their life in general. The paper will first introduce major tests and examinations of English designed and administered in China, then provide an overview of the current research in language testing that has been conducted by Chinese researchers and published in Chinese academic journals over the past 10 years. This paper will focus on the discussion of the issues and concerns of language testing within the Chinese context.

I Introduction

The People's Republic of China (China, 中國: ***Zhōngguó***[1]) is a vast geographical region of about 9.6 million square kilometers with over one-fifth of the world's population (1.3 billion in 2005). The majority of China exists today as a country known as the People's Republic of China, but it also refers to a long-standing civilization comprising successive states and cultures dating back more than 4000 years.

The testing and examination history in China can be traced back to the imperial period nearly two thousand years ago since the Han Dynasty (206 BC to AD 220). The imperial examinations[2] (科舉: ***kējǔ***) in dynastic China is the first of its kind used to select the highest officials

Address for correspondence: Liying Cheng, Faculty of Education, Queen's University, Kingston, Ontario K7P 3N6, Canada; email: chengl@educ.queensu.ca

[1] The phonetic symbol (pinyin) of the Chinese characters is in italics throughout this paper.
[2] The content and method of the examinations changed over time. '6 Arts' – music, archery, horsemanship, writing, arithmetic, and ceremonial rites – were tested in 1100 BC. '5 Studies' – civil and military law, agriulture, taxation, and geography – were tested in 202 BC – AD 200. By the seventh century AD, the examinations became a national system by which candidates' abikity to remember and interpret Confucian classics was measured (see Higgins and Sun, 2002; Zhang, 1988).

 DOI:10.1177/0265532207083743

of the country. These examinations determined the positions in the civil service based on merit and education, which promoted upward mobility among the population for centuries. In addition, these examinations are regarded by most historians as the first standardized tests based on merit (Hu, 1984; Lai, 1970) and offer the biggest contribution that ancient China made to the testing movement (Higgins and Sun, 2002). This tradition of using examinations for selection is still evident in the current education system in China. A student can start to take examinations as early as the entrance test to enter kindergarten at the age of four. Over the years of primary education (K–Grade 6), secondary education (Junior High Grade 7–9, Senior High Grade 10–12) and university education (4-year undergraduate), students take numerous examinations at the school, municipal, provincial and national levels. Furthermore, examinations continue to enjoy a wide societal acceptance and recognition in China as fair measurement for selection of the best talent into the social hierarchy (Cheng and Qi, 2006).

However, the testing of foreign languages started much later. In 1862, the first school of foreign languages, *Jing-shi-tong-wen-guan* (Beijing Normal Language School), was established in China (Ke, 1986) with British missionary J.C. Burdon as its first English instructor. During its 40-year history, the Beijing Normal Language School taught only five foreign languages: English, French, Russian, German and Japanese. In 1901, the school merged with *Jing-shi-da-xue-tang* (Beijing Normal University), established in 1898, and was renamed Peking University in 1912 (He, 2001). The early schools of foreign languages were often small in scale and aimed at training diplomats and translators for the Chinese government. Later on, more schools were set up, where foreign languages were taught and assessed, signifying the beginning of foreign language testing in China.

The economic reforms in China in the late 1970s and the opening of China to the outside world have brought about rapid and drastic changes in China over the last 30 years. During the same period, there has also been a great boom in foreign language education in China. Foreign language education, especially English education, has become more and more important for Chinese learners at all levels of education. Since the mid-1990s, English began to be taught from Grade 3 in primary education. English, as one of the three core subjects along with mathematics and Chinese, is tested for students to enter a junior and senior high school. English is a compulsory subject in the national university entrance examinations for all types of universities and colleges. English is also an obligatory subject for all majors in Chinese universities and colleges. Non-English majors are

required to take the college English course for at least two years. To obtain a bachelor's degree in Chinese universities, these students often need to pass the College English Test – an English language proficiency test. English is an examination subject for all students who wish to pursue a graduate degree in China. Apart from English as an academic requirement, English skills are tested for all those seeking promotion in governmental, educational, scientific research, medical, financial, business and other government-supported institutions (He, 2001). Therefore, it is no exaggeration to say that China has the largest English-learning population in the world. In reality, being successful in the various English tests and examinations is the key to the success in life for many in China and also for many Chinese who wish to pursue their undergraduate and graduate education in English-speaking countries (including the possibility of emigration).

As can be seen from the length of its history and the huge population of English learners and test-takers in China, it is simply not possible to discuss in depth the issue of English testing in a paper of this scope. The paper will thus first introduce major tests and examinations of English designed and administered in China, then provide an overview the current research in language testing that has been conducted by Chinese researchers and (for the most part) published in 10 key academic journals in the area of foreign language teaching and learning in China from the year 1995 to 2006.[3] These journals are published in Chinese.[4] This paper thus serves to introduce the issues and concerns of language testing within the Chinese context to a wider audience.

II English test development in China

Now in China, major English tests designed locally include the College English Test (CET), the National Matriculation English Test (NMET), the Test for English Majors (TEM), the Graduate School Entrance English Examination (GSEEE) (研究生英语入学考试), the

---

[3] The 10 Chinese academic journals cited in this paper are: *Foreign Language Education, Foreign Language Teaching and Research, Foreign Language Teaching and Research in Basic Education, Foreign Language World, Foreign Languages and Their Teaching, Hubei Zhaosheng Kaoshi (entrance examination), Journal of PLA University of Foreign Languages, Journal of Tianjin University of Commerce, Jouranl of Xi'an Foreign Languages, Modern Foreign Languages,* and *Teaching English in China*.

[4] The only exception is the journal *Teaching English in China*, which is published in English.

Public English Testing System (PETS), the Cambridge Business English Certificate (BEC), and the WSK – an examination to select professionals for study and training overseas.

The College English Test (CET) is a large-scale standardized test administered nationwide by the National College English Testing Committee on behalf of the Higher Education Department of the Ministry of Education (MoE) in China. It aims at measuring the English proficiency of college/university undergraduate students in accordance with the College English Teaching Syllabus (*National College English Syllabus for Non-English Majors*, 1999). The CET is a test battery, which is comprised of the CET Band 4 (CET-4), the CET Band 6 (CET-6), and the CET – Spoken English Test (CET-SET). The CET-4 and CET-6 were first administered in 1987, and are administered twice a year afterwards, in January and June. The CET-4 and CET-6 are criterion-related norm-referenced tests (Jin, 2005; Yang and Weir, 2001). In this sense, the test criteria are based on the National College English Syllabus by the Higher Education Department of the Ministry of Education to guide the English teaching at university level. The test scores are based on 100 points, of which a score of 60 points constitutes a passing grade. The reported score of the CET conveys two pieces of information. First, it indicates whether a candidate has met the requirements of the College English Teaching Syllabus (*National College English Syllabus for Non-English Majors*, 1999). Second,, it indicates the percentile position of a candidate in the norm group, which consists of over 10 000 college/ university students from six top universities in China.

In order to meet the needs of China's economic reform and open-door policy, the CET-SET was included in the test battery in 1999. The CET-SET is available to students who have passed the CET-4 with a score of 80 or above or the CET-6 with a score of 75 or above. The CET-6 is available to students who have passed the CET-4, and have taken the optional College English Course of Bands 5–6. Certificates are issued by the Higher Education Department of the Ministry of Education, P. R. China, to those who qualify on the test. Since its inception in 1987, the CET has attracted an increasing number of candidates every year. In the 2005 academic year, more than 9.58 million students in China took the test (Jin, 2005). The CET is reported to have maintained high reliability and validity (Jin, 2005; Yang and Weir, 2001) and has a set of standardized procedures in the administration and the interpretation of raw scores (Yang and Jin, 2000).

Consequently, the stakes associated with the CET are extremely high. In most colleges and universities, the CET-4 certificate is one

of the graduation requirements for students to obtain their academic degree. The CET certificate is also an asset for university graduates who want to stand a better chance in the job market. Students' performance in the CET-4 also affects the evaluation of teachers, their promotions, and even merit awards. At the institution level, the passing rate on the CET is often regarded as one of the criteria to judge the prestige of a university. Therefore, the CET has exerted a huge amount of influence, reportedly negative, on English language teaching and learning at the tertiary level in China since its first administration (Han *et al.*, 2004; Zhang, 2003). Many language educators and researchers have started to investigate various aspects of the test, hoping to bring some insights into the improvement and innovation of the test. This will be discussed in more detail later.

The National Matriculation English Test (NMET) is the university entrance test of English for the whole country. The purpose of the test is to make inferences about candidates' English language ability, which is used in university admission decisions together with the scores from university entrance tests in another five or six secondary school subjects. A student needs to take tests in five or six subjects depending on the requirements of the type of the university for which he/she applies. Chinese, mathematics, and English are three compulsory subjects for all candidates regardless of their choice of university. The NMET is taken annually on 8 June by millions of secondary school graduates who wish to gain entrance to Chinese universities and colleges. The number of candidates varies each year. It was about three million in the early 1980s and increased to 9.5 million in 2006 (http://www.sina.com.cn). The test is about two hours long. It is administered by the municipal, county, and provincial Bureau of Education at testing centres across the entire country (Cheng and Qi, 2006). The NMET, introduced in 1985, is one of the three compulsory tests in the university entrance examination battery in China. It is a norm-referenced standardized test with a major function of selecting secondary school graduates for institutions of higher education. The NMET is a high-stakes English test in China, which influences how society evaluates the schools, and how schools, parents, and students evaluate the teachers. Apart from its selection purpose, the NMET was used to bring positive washback to English teaching and learning at the secondary level (Qi, 2003; 2004; 2005).

For English majors in Chinese colleges and universities, the Test for English Majors (TEM) is an important test. The TEM assesses the language performance of English majors and is administrated by the National Advisory Commission on Foreign Language Teaching in

Higher Education (NACFLT) in China. Another purpose of the test is to promote English teaching and learning for English majors. The TEM is a criterion-referenced test. Students' performance is evaluated against the criteria stipulated by the teaching syllabus (Zou, 2003). The test consists of two levels: TEM-4 administered at the end of the 2nd year, and TEM-8 at the end of the 4th year in their undergraduate program. The Graduate School Entrance English Examination (GSEEE) is another entry test administered at the national level once every year for entrance into graduate schools at all Chinese universities. The GSEEE is administered by the National Education Examinations Authority of the Ministry of Education in China.

Apart from these tests in institutional settings, there are some non-credential English tests in China. The Public English Testing System (PETS) is probably the largest in scale among them. It was developed in 1999 by the Chinese National Education Examinations Authority (NEEA) with assistance from the University of Cambridge Local Examinations Syndicate (UCLES). This test is a non-credential test, which is open to all English learners, with no restriction on age, profession or academic background. It aims to promote English learning nationwide. It provides assessment and certification of communicative English language skills in reading, writing, listening, and speaking at five levels of competence from Level 1 to Level 5. Another public test is the Cambridge Business English Certificate (BEC), another collaborative program between the NEEA and UCLES, designed to test English language ability used in the business context. The test was introduced to China in 1993 with three levels – BEC Preliminary, BEC Vantage, and BEC Higher. Individual learners who wish to obtain a business-related English language qualification sit for this test. BEC certificate is widely recognized by foreign companies and enterprises in China.

In addition, every year, the Chinese government provides funding for professionals to study and/or receive training outside China. Apart from their professional qualifications, these professionals (non-foreign language majors) are selected based on a language proficiency test – WSK (an acronym from Chinese pinyin: *wai-yu-shui-ping-kao-shi* 外语水平考试) administered by the National Education Examinations Authority (NEEA) of Ministry of Education. WSK provides tests of communicative competence in five languages: English, French, German, Japanese, and Russian.

Compared with the academic tests, the non-credential tests of English receive less attention from language educators and researchers. Many more empirical studies are found in academic journals

investigating tests such as the CET and the NMET – the two large-scale high-stakes national tests.

III Empirical studies in language testing

As mentioned above, examinations have played an important social and educational role in China. The promotion of an effective English testing system has thus been of great importance in China. Particularly for the past 15 years, language educators and researchers in China have devoted much effort to issues in language testing. Due to the research tradition in China with its focus on knowledge dissemination, a fair number of the published articles on language testing in Chinese academic journals are review articles or state-of-the-art articles synthesized by known researchers in the field of language testing in China. These publications offer an insight into the kind of discussions on language testing and the theories and models of language testing that have been introduced to Chinese academics and researchers (see Han, 1995, 2003). However, due to the space limitation, I have only reported empirical studies here. A fair number of empirical studies have been conducted in China and published in Chinese academic journals. Only a few have been published in academic journals outside China. These studies are primarily in the areas of test validity, testing of speaking, test-taker characteristics; and test washback, which represent the main issues in language testing within the Chinese context.

1 Test validity

The issue of test validity has drawn a fair amount of attention from Chinese language testing researchers. Much has been discussed about the validity of a widely used test format in China – multiple-choice questions. Considering the huge population of millions of test-takers each year in China, many of the language tests – designed and administered in China as mentioned above – adopt large numbers of multiple-choice questions in order to save time and manpower that would otherwise be involved in grading the test papers. Sun (2000) conducted a study to evaluate the test items in a language proficiency test for specific purposes. Each test item was examined in three aspects: item difficulty, passing rate, and discrimination. The results were used to validate an item, and to determine whether the distractors for each item are overly distracting or whether the test-takers had the required

language knowledge. This study has implications for both test constructors as a basis for improving the items, and for teachers to draw attention to aspects of language knowledge that may have been neglected in their teaching.

Wang (1996) conducted a comparative study of multiple-choice and true–false questions to evaluate the effectiveness of the latter as a test format. A test with 60 multiple-choice questions and a test with 60 true–false questions were administered to three groups of around 30 students representing different levels of language proficiency. It was shown that the efficiency of true–false test was 1.18 times greater than that of the multiple-choice tests and that language proficiency level did not influence the time taken to complete the test. However, both the reliability and discrimination indices of the true–false test were lower than those of the multiple-choice test. The researcher concluded that the true–false test format can be used for classroom assessment, but not in a large-scale standardized test.

It is commonly believed among teachers and students in China that students do not need to read carefully or even comprehend passages to pass a test, and that multiple-choice reading comprehension tests do not accurately indicate students' actual reading comprehension ability. Based on this assumption, Cheng and Gao (2002) explored the extent to which Chinese university students rely on reading passages in answering multiple-choice reading comprehension tests. This study examined the reading test scores of groups of Chinese university students taking the standardized multiple-choice reading comprehension test of the College English Test in China under two different sets of testing conditions. Sample CET test papers were used for this study. Under the first testing condition, the test performance was compared between one group of students who were allowed to do the test as in a normal reading comprehension testing situation and another which was allowed to read the passages only once and then do the multiple-choice questions without going back and forth with the reading passages. The second condition compared the test performance of two groups of students in a with- or without-passage situation. The findings show firstly that students perform better when they are not allowed to go back and forth between the questions and passages when compared with a normal reading comprehension test situation and second, that passage comprehension is relevant to reading comprehension test performance (i.e. students need to read the passages in order to answer most questions correctly). However, in the second condition, even without the associated passages, students still achieved scores above chance

level (39.91% and 28.89% respectively on MC items with four choices) with some test items showing a mean item difficulty of 0.94 and 0.88. This suggested that the guessing was prevalent among the participants in this study.

Zhou (2004) conducted a comparability study of two national tests – CET-6 (for non-English majors) and TEM-4 (for English majors) in terms of test-takers, test scores, and test content. The results show modest covariance of the two tests with a Pearson correlation coefficient of 0.712 ($p < 0.01$). The two types of test-takers were quite similar to each other, with the only obvious difference being in the hours of English instruction they had received. The two tests are related, both in their power to measure language ability, and in terms of test methods. The researcher concluded that either test could serve the purpose of measuring language proficiency. This study provided insights and posed questions about the validity of the two big English tests in China.

2 Testing speaking

The development of oral testing only started in the 1990s in China, where the tradition of language testing had solely focused on reading and writing. The Cambridge Business English Certificate (BEC) was introduced to China in 1993 with an oral component. The oral test in the Test for English Majors (TEM) battery started in 1994. The National Matriculation English Test – Oral Subtest (NMETOS) was formally introduced in 1995 in three provinces in China (see Li and Wang, 2000). The College English Test – Spoken English Test (CET-SET) was introduced in 1999 (see Jin, 2000b; Huang, 1999; Yang, 1999). Language testing researchers in China have conducted a number of empirical studies to investigate various issues concerning oral test development.

Li and Wang (2000) reported on the development of the National Matriculation English Test – oral subtest. They discussed the many limitations including the huge size of the test candidature and the rigid limitation on human and time resources in China. Nevertheless the validity of the NMETOS format was asserted on the strength of its being a message-based test of spoken interaction achieving a balance between control and spontaneity in the required spoken output, a union of the analytical and the holistic approach to rating, and a combination of the single-examiner method in test administration and a double-marking method for scoring. The authors argued that 'the NMETOS is not just a test that suits the conditions and meets the needs of China. In the wider context of language testing, it can claim to be a successful innovation in mass-scale oral testing' (Li and Wang, 2000: 160).

Li (1999) reported on the development of the CET-Spoken English Test, which was based on Bachman and Palmer's (1996) theoretical framework of language test design. The first task of the test development was to specify the target language use domain and language use tasks for the test. A survey was conducted of the needs for speaking abilities in different work settings, with an attempt to find out the future real-life language use domain for target test-takers. The test developers realized that the survey, because of its scale, was far from being complete in terms of specifying all the possible future language use domains for college/university students in China, and that this might result in bias in test task design against certain sub-groups of the test-taker population. In addition, the requirements of the College English Teaching Syllabus with regard to speaking abilities were also taken into consideration. To choose suitable task types for the test as the next step in test development, another two surveys were administered to college/university students and their English teachers. These surveys explored the current speaking abilities of the students, and the prevalent oral activities in their classrooms. It also investigated characteristics of Chinese college/university students, such as age, educational background, aptitude, and attitudes towards speaking tests, as it was believed that these might influence test performance. The researcher stated that the development of the test was never an easy task and further research is needed. It is a daunting task to find an effective way to assess the speaking ability of a large population of non-English major students in China. This is one of the reasons that currently CET-SET is only accessible to a smaller number of students (those who passed the CET-4 with a score of 80 or above or the CET-6 with a score of 75 or above) given the time and cost involved in the test administration and rater training.

He and Dai (2006) conducted a corpus-based investigation into the validity of the CET-SET. They examined the degree of interaction among candidates in the group discussion task with respect to a set of interactional language functions (ILFs) to be assessed. Their results showed a low degree of interaction among candidates in this task. The researchers discussed a variety of factors that may explain the low degree of interaction and they suggest that 'the inadequate elicitation of ILFs from the candidates may well pose a problem for measuring their speaking ability in terms of the ability to engage in communicative interaction' (p. 393).

In an attempt to find an alternative way of conducting a large-scale speaking test, Xiong, Chen, Liu, and Huang (2002) carried out an experimental study on a semi-direct oral test, or 'Recording Oral

Test' as the researchers called it (see Xiong *et al*., 2002: 283 – 錄音口語測試). In the experiment, the test takers were asked to speak into a microphone after being given a prompt from the tape, rather than in front of an interlocutor in a face-to-face oral test. The study involved the design of test content and rating scales. Three different analytic rating scales were used to evaluate each student's performance: an ability scale, an item scale, and a holistic scale. The purpose was to ensure the reliability of the test score. The data analysis showed a high correlation among the scores obtained from the three scales. A high correlation was also reported between student's ranking in class given by their classroom teachers and each of the three scores, which was interpreted as evidence that the students have demonstrated their language abilities in the Recording Oral Test. The researchers concluded that conducting a recording oral test was feasible as an alternative way to assess speaking abilities. However, it is still too early to jump to the conclusion that a semi-direct oral test as such can prove able to replace a face-to-face interview. Actually, growing evidence has been gathered which favors direct over semi-direct tests in terms of validity (Fulcher, 2003). Studies by Shohamy (1994) and O'Loughlin (2001) also suggest that direct tests and semi-direct tests measure different constructs. Therefore, in situations where practical considerations do not allow a direct test to be used, it is necessary and important to revisit the construct definition of a semi-direct rest so that the test score interpretation does not go beyond the test constructs (Fulcher, 2003). Other factors also need to be addressed in relation to the EFL context in China such as variation in discourse across the direct and semi-direct tests, frequency of errors and pauses, and degree of test anxiety in the two different testing situations.

Guo (1999) also provided some relevant information in the above areas, conducting a situational difference test of a group of final year English majors to explore how changes in situation would influence oral language performance. The students were tested in three situations: (a) recording their opinions of a topic on a tape (S1), i.e. they talked to a tape-recorder; (b) talking to some freshmen in a casual environment on the same topic (S2); and (c) talking to a tester in an office, again on the same topic (S3). The students were also required to complete a motivation questionnaire. The purpose was to explore the correlation between their motivation and oral performance in each situation. Students' performance was evaluated in two areas: length of natural pauses and frequency of unnatural pauses. Results showed a high correlation between students' motivation and the length and frequency of pauses in both S1 and S3, but not in S2. The researcher

inferred that the pressure students felt in completing a task varied in the three different situations, causing different degrees of anxiety, with consequences for their language fluency. In a casual environment as in S2, the fewest unnatural pauses occurred. While the generalizability of this research is limited by the fact that there were only 10 participants involved in the study, the researcher suggested that test-takers' affective factors involved in different situations should be considered in the development of oral language tests.

3 Test-taker characteristics

Language testing researchers in China have consequently investigated the relationship between test-taker characteristics and language test performance. Zeng (2002a) conducted a study to explore the relationship between one personal characteristic (self-confidence) and test performance. In his study, 170 students participated in a computer-based test. During the test, the students were asked to evaluate their self-confidence in answering each item on a 7-point Likert scale. The students' test score, average time spent on each item, and scores of self-confidence were then calculated. The scores for self-confidence included the mean score for self-confidence, right answer self-confidence, wrong answer self-confidence, and a self-confidence difference index (i.e. the difference between right and wrong answer confidence). The item-level difficulty of the test was also measured with the use of Gitest 2.0.[5] Results showed a high correlation between students' confidence score and the average time spent on each item, and also between confidence score and item difficulty, indicating that students with a high level of confidence spent less time in answering each item, and suggesting that the item difficulty influenced their level of confidence. A correlation of 0.554 ($p < 0.01$) was observed between the students' test score and the mean confidence score, and a correlation of 0.416 ($p < 0.01$) between test score and right answer confidence, while a correlation of -0.91 existed between test score and wrong answer confidence. This indicated that students with a high test score were more confident. The researcher also found four patterns of relationship between test performance and level of confidence: (a) right answer with high level of confidence (P1); (b) right answer with low level of confidence (P2); wrong answer with high

[5] Gitest is a test analysis system developed by Professor Shichun Gui and Li Wei from Guangdong University of Foreign Studies, Guangzhou, China.

level of confidence (P3); and (d) wrong answer with low confidence (P4). The following inferences were drawn by the researcher. P1 indicates that the students' demonstrated their real language ability in the test. P2 means there might be guessing involved in getting the right answer. P4 shows that the students did not acquire the knowledge they were supposed to. The situation of P3 is more complicated. Many factors may contribute to the occurrence of this pattern, as the researcher explained. For example, the students might be overconfident, or some knowledge might cause them confusion, or there could be problems with the test item itself, which needs to be investigated further. Research findings in this area should be beneficial both for the improvement of a test and for the promotion of language teaching and learning.

Based on this study, Zeng (2002b) proposed a model for an individualized self-adaptive test design, in which self-confidence is considered. According to this model, students first evaluate their own language ability. Based on their self-assessment, the first test item will be chosen for each individual test-taker by the computer. After answering each question, students will be asked to score their confidence in choosing the answer. Then students' language ability will be re-estimated. According to the item difficulty and the student's confidence score, the relative difficulty of that item for the student will be calculated. When two different students choose the same answer, if their score of confidence is different, the relative difficulty will be different, which means their ability estimate will be different. Based on this estimation, the difficulty of the next test item is adjusted and chosen for each individual test-taker. This adaptive process continues until an acceptable degree of test accuracy is achieved.

The researcher used this model to design an individualized self-adaptive test (ISAT) and administered it to a group of 112 students, together with a computer adaptive test (CAT), and a self-adaptive test (SAT). The three tests were at the same level of difficulty. The purpose of the study was to test the effectiveness of ISAT and to explore its advantages over CAT and SAT. Data analysis first involved a comparison of the three tests in terms of students' mean score ability estimate, standard deviation, number of test items needed, test accuracy, and the average time spent on each item. Results showed that for ISAT, a smaller number of test items were needed than was the case for the other two tests to reach the same level of test accuracy. Also, the average time the students spent in answering an item was 3 to 5 seconds shorter than in ISAT. A correlation was also reported between the three tests, and between the students' score ranking in

ISAT and their ranking in class given by their teachers, which showed that ISAT was an effective way to assess the students' language ability. The researcher made the assumption that including students' confidence score as a criterion for adjustment would enhance the level of adaptability of the test. This could further differentiate the ability of two students who choose the same answer but with a different level of confidence, an advantage, as the researcher believed, over the maximum likelihood estimate in IRT, which would identify the same two students as having the same level of ability. Thus, it is argued, the ISAT caters more to test-takers' individuality. Zeng's studies have investigated the relationship between the other personal characteristics and language test performance in the Chinese EFL context and have offered insights into the development of computer-based language tests in China.

Song and Cheng (2006) examined language learner strategy use reported by 121 Chinese learners of English through a questionnaire and the relationships between their strategy use and language performance on the College English Test Band 4 (CET-4). Results showed that the participants of the study reported using more metacognitive strategies than cognitive strategies in general. One subscale of cognitive strategies – *inferencing* – however, was reportedly used most frequently. *Memory and retrieval strategies* – as a subscale of cognitive strategies in the questionnaire – were the only significant predictor of the CET-4, accounting for 8.6% of the CET-4 variance. The complex relationships between learner strategy use and their test performance are discussed in detail. In a similar study, Zhang (2004) investigated the relative contributions of cognitive and metacognitive strategies, deep and surface approaches, grammar, and vocabulary to reading performance within the context of the College English Test, and the interrelationships among these variables. Participants were 435 first-year students at an urban university in Beijing, China, who completed 'the approaches to learning questionnaire, the cognitive and metacognitive strategies questionnaire, a vocabulary test, and two grammar and reading tests' (p. iv). Exploratory factor analysis identified two new constructs: metacognitive awareness and test-taking strategies. The results of regression analyses and structural equation modeling, however, showed that vocabulary and grammar made the greatest contribution to EFL reading performance. Test-taking strategies and the deep and surface approaches (as defined by the approaches to learning questionnaire) did not predict reading performance. Metacognitive awareness made a

significant contribution to reading only in the absence of vocabulary and grammar, and the effects were weak. The deep approach was associated with metacognitive awareness. Test-taking strategies were correlated with both the deep and surface approaches. Test-taking strategies had a negative direct effect on grammar, and a negative indirect effect on reading via grammar. Metacognitive awareness had both negative and positive effects on reading. The implications for practice are that basic skills need to be strengthened in addition to (or before) attempting to strengthen cognitive and metacognitive strategies. In the EFL teaching environment, teachers should pay attention to enhancing students' language proficiency before strategy instruction intervention.

4 Test washback

The tradition of using examinations for the purpose of selection in China has put testing in a position to affect the huge number of stakeholders involved. This influence of testing has long been discussed, especially in relation to nationwide tests, for example, the College English Test (CET) and the National Matriculation English Test (NMET). Interestingly, both tests serve more than one function in Chinese society. The CET is a norm-referenced test, but also criterion-related, in that it aims to measure the English proficiency of college/university undergraduate students in accordance with the College English Teaching Syllabus. The NMET, apart from its primary function of selecting candidates for institutions of higher education, is designed specifically to promote changes in English language teaching in schools.

Qi (2003; 2004; 2005) investigated the intended washback of the National Matriculation English Test in China (NMET). This study examined the reasons why the NMET failed to bring about the intended changes or washback effects on the teaching and learning of English in secondary schools. For this purpose, data were collected through interview and questionnaire from eight NMET constructors, six English inspectors, 388 teachers and 986 students. The results show that the most important reason for the test failing to achieve the intended washback is that its two major functions – the selection function and the function of promoting change – are in many ways in conflict with each other, making it a powerful trigger for teaching to the test but an ineffective agent for changing teaching and learning in the way intended by its constructors and the policymakers. Analyses of interviews data revealed that there was considerable

discrepancy between the test constructors' intentions and school practice. The study concluded that the NMET has achieved very limited intended washback and the test is an inefficient tool for bringing about pedagogical changes in schools in China.

Gu (2005) explored the relationship between the CET and college English (CE) teaching and learning. The research focused on: (1) the CET participants' perceptions of the test and its washback; (2) the processes of CE classroom teaching and learning, including CET washback on CE classroom teaching and learning; and (3) the products of CE teaching and learning. In addition, other major factors exerting influence on CE teaching and learning were analyzed. The study was carried out in both case study settings and nationwide contexts. A wide range of CET stakeholders (e.g. administrators, teachers, and students), about 4500 in total, were involved. Various research methods were employed, including classroom observations, questionnaire surveys, interviews, tests and analyses of documents, of 'coaching materials', as well as of CET data and of the examinee output in the CET. The findings showed both positive and negative washback of the CET. Most of the CET stakeholders think highly of the test, especially its design, administration, marking and the new measures adopted in recent years. They believe that the positive washback of the test is much greater than the negative washback, and that the negative washback is primarily due to the misuse of the test. However, some CET stakeholders are dissatisfied with the overuse of the multiple-choice (MC) format in the test, the lack of direct score reports to the teachers, the incomplete evaluation of the students' English proficiency without a compulsory spoken English test, and the use of the test as the sole means in evaluating the quality of CE teaching and learning. The study concluded that the issue of the CET washback is complicated and pointed out that the CET is part of a complex set of factors that determine the outcome of CE teaching and learning. The top three factors within the school context are: students' educational background, teacher quality, and administrators' attitudes about the CE courses and the CET.

Han, Dai, and Yang (2004) conducted a survey among 1194 English teachers in 40 colleges and universities, asking about their attitudes toward the national testing system of the CET at the tertiary level. They found that 37.7% of the teachers thought that the CET pushed colleges and universities to use the passing rate of the test to evaluate their teaching. Over 70% of the teachers did not believe that the test could improve overall English teaching and learning at the tertiary level in China. About 25% of the teachers pointed out that

the test encouraged students to guess and to use test-taking strategies, rather than to improve their actual language ability, and 37.8% of the teachers attributed the lack of communicative competence of their students to this test. However, about 70% of the teachers did not want the test to be abolished. From the interviews with some university administrators and English teachers, the researchers found that one reason for this contradiction in attitudes was the time and effort that would have been consumed to design their own test systems and to grade large numbers of test papers. Another concern was the validity issue of a possible self-designed test by an individual university. In terms of classroom teaching, about 40% of the teachers believed that the CET influenced regular teaching. When asked about a suitable type of a national test for college English teaching, 40% of the teachers thought that it should be a language proficiency test rather than an achievement test, and 45.4% of the teachers suggested that all four skills should be assessed in order to promote students' overall language competence. The teachers were also asked their opinions regarding the relationship between the CET certificate and students' actual language ability. Most of the teachers (77.9%) did not think that these two components were correlated, i.e. having a CET certificate does not necessarily mean that the student has the language competence as required by the College English Syllabus. These findings showed that teachers were doubtful about the validity of the CET.

Jin (2000a; 2000b) examined the washback effects of the CET-Spoken English Test. Questionnaires were distributed to 358 students who took the test in the year of 1999, and to 28 English teachers who worked as interviewers in the test. The questionnaire covered the following areas: students' motivation to take the test, the importance of the test, and its potential washback effects. A large number of students (79.6%) reported that they took the test to have their communicative competence in English evaluated. Most of the students (96.9%) and teachers (100%) thought that it was important to have an oral test in the CET battery. All of the teachers believed that the Spoken English Test would have a huge impact on college English teaching and would promote students' ability to use English communicatively; 92.3% of the students and all the teachers suggested that the test should be accessible to a larger number of students. The questionnaire also asked the teachers and the students to evaluate the test design, which included test method, test format, test tasks, test time, the reliability of the test, and the rating scale. The results were very positive. The researcher claimed that since the administration of the CET-SET, positive changes have taken place in college English teaching. For example, many

colleges and universities began to pay more attention to improving students' communicative competence; students became more involved in the oral activities in class; and some universities even developed teaching materials that catered to the test. However, there is a lack of empirical studies or evidence to support these claims so far.

Zou (2003) investigated the mutual influence of the Test for English Majors (TEM-8) and the teaching syllabus for English majors, by revisiting the TEM-8 test development and principle of test design, and by analyzing the teaching syllabus. She believed that the teaching syllabus guided the TEM-8 test development, and the feedback from the test results contributed to the improvement and revision of the teaching syllabus.

The above literature addressed the washback of the NMET, the CET and TEM-8 on teaching and learning without linking impact to actual test performance. Zhao (2006) goes a step further in exploring Chinese university students' attitudes toward the CET as well as the relationship between their attitudes and their test performance. The results of the study showed that students held strong yet mixed feelings toward the CET-4. On the one hand, they were motivated to do well on the CET-4; on the other hand, they were not sure of their ability to perform well on the test. Two factors, test-taking motivation and test-taking anxiety/lack of concentration, were the best predictors of students' test performance on the CET-4. Student's attitudes toward the CET-4 accounted for about 15.4% of the variance in their test performance. The factor test-taking anxiety/lack of concentration also differentiated female and male students. Three factors – test-taking anxiety/lack of concentration, test-taking motivation, and belief in CET-4 – differentiated high-achieving students from low-achieving students.

The administration of these large-scale standardized language tests has been a controversial issue in China. For example, the CET is a test that has been more frequently researched over the past ten years. Some researchers point out the positive impact of the CET as it promotes the role of English teaching and learning at the tertiary level in China (see, for example, Li, 2002). Other researchers (e.g. Han *et al*., 2004; Gu and Liu, 2002) challenged the validity of the CET by pointing out that the test does not assess communicative competence as the teaching syllabus requires. Many researchers have advocated that testing should support teaching while in reality the CET drives teaching in China. It is not an exaggeration to say that

the CET is probably the most debated test in the language testing field and among academics in China.

IV Conclusion

The combination of a very long history of using tests and examinations for selection purposes in Chinese society and the relatively new development of English testing in China (the NMET introduced in 1985 and the CET in 1987) has provided unique challenges to English language education in China. On the one hand, the Chinese society in general accepts the function of testing as a fair indicator of students' academic success. Consequently, teachers and students follow the testing in their teaching and learning and/or make passing the test the goal of their teaching and learning. On the other hand, it is widely acknowledged that the English testing system (as well as English teaching and learning) needs to be further enhanced in terms of its quality. More empirical research needs to be conducted to provide evidence for the validity of the tests. The impact of the two national tests – the NMET and the CET – is in particular evident on the teaching and learning of English in China and has been much debated.

As a result of these debates, more recent publications have focused on investigating the relationship of testing and teaching and the factors contributing to the test score. For example, Kang and Chen (2005) revisited the nature of the CET-4, taking into account the ethical use of the test, its design principle, its scoring system and the test-taking process. The authors propose that the CET should be better defined as a norm-related criterion-referenced achievement test to achieve a positive impact educationally, socially and ethically. Pan (1998) critiqued the CET-4 in relation to the college English teaching by looking at the achievements and shortcomings of the test (and test formats) and suggested further reform of the college English teaching and testing system. Given that the college English teaching is shifting to a communicative approach, Wang and Zhou (2005) conducted a validation study of the CET and proposed an alternative university-based communication test. Lu (2005) investigated the factors which influenced reading comprehension results between those who passed and failed CET-6. The author concluded that extra reading is closely related with the students' reading score and that teachers should play an active role in supporting students' learning in this area.

In addition, a number of research studies started to shift the focus on testing to assessment within the context of classroom teaching. For example, Yu (2005) emphasized the importance of linking instruction and assessment in English teaching by presenting alternative assessment approaches in the classroom. It is evident that researchers in China are making greater efforts to understand the relationship between testing, and teaching and learning of English. The key to success for Chinese students should not simply be successfully passing an English test, but to become a fluent user of English in their academic study and in their future workplace.

Acknowledgements

The author would like to thank Ying Yu and Jing Zhao for their support in researching and accessing research studies published in the Chinese academic journals cited in this article while they were pursuing their Master's of Education degrees at Queen's University, Kingston, Canada. This article was supported by an MOE Project of the National Key Centre for Linguistics and Applied Linguistics of Guangdong University of Foreign Studies, China.

V References

Bachman, L.F. and **Palmer, A.S.** 1996: *Language testing in practice*. Oxford: Oxford University Press.

Cheng, L. and **Gao, L.** 2002: Passage dependence in standardized reading comprehension: Exploring the College English Test. *Asian Journal of English Language Teaching* 12: 161–78.

Cheng, L. and **Qi, L.** 2006: Description and examination of the National Matriculation English Test in China. *Language Assessment Quarterly: An International Journal* 3(1): 53–70.

Fulcher, G. 2003: *Testing second language speaking*. London: Pearson Education.

Gu, X. 2005: Positive or negative? An empirical study of CET washback on college English teaching and learning in China. *ILTA Online Newsletter*, 2. http://www.iltaonline.com/newsletter/02–2005oct/ (last accessed 1 June 2006).

Gu, W. and **Liu, J.** 2005: Test Analysis of College Students Communicative Competence in English. *Asian EFL Journal* 7(2): 118–33.

Guo, Q. 1999: The influence of language environment on language output. *Foreign Language Teaching and Research* 31(1): 35–40.

Han, B. 1995: Lyle F. Bachman's theoretical model for language testing. *Foreign Language Teaching and Research* 27(2): 55–60.

—— 2000: Language testing: Theories, practice, and development. *Foreign Language Teaching and Research* 32(1): 47–52.

—— 2003: New development in language testing: Task-based language assessment. *Foreign Language Teaching and Research* 35(5): 352–58.

Han, B., Dai, M. and **Yang, L.** 2004: Problems with College English Test as emerged from a survey. *Foreign Languages and Their Teaching* 179(2): 17–23.

He, L. and **Dai, Y.** 2006: A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing* 23(3): 370–401.

He, Q. 2001: English language education in China. In Baker, S.J., editor, *Language policy: Lessons from global models*. Monterey, CA: Monterey Institute of International Studies, 225–31.

Higgins, L. and **Sun, C.H.** 2002: The development of psychological testing in China. *International Journal of Psychology* 37(4): 246–54.

Hu, C.T. 1984: The historical background: Examinations and controls in premodern China. *Comparative Education* 20: 7–26.

Huang, P. 1999: The College English Test-Spoken English Test development and college English teaching. *Foreign Languages and Their Teaching* 118(3): 21–23.

Jin, Y. 2000a: The washback effects of College English Test-Spoken English Test on teaching. *Foreign Language World* 118(2): 56–61.

—— 2000b: Feedback on the CET Spoken English Test and its backwash effect on the teaching of oral English in China. *Proceedings at the Third International Conference on English Language Testing in Asia*. Hong Kong: Hong Kong Examinations Authority, 205–14.

—— 2005, August: *The National College English Test of China*. In Hamp-Lyons, L. (Chair), *The big tests: Intentions and evidence*. Symposium presented at International Association of Applied Linguistics (AILA) 2005 Conference in Madison, WI.

Kang, Y. and **Chen, J.** 2005: Testing the test: Aspects of CET revisited. *Teaching English in China* 28(2): 21–25.

Ke, F. 1986: *History of Foreign Language Education in China*. Shanghai: Shanghai Foreign Language Education Press.

Lai, C. T. 1970: *A scholar in imperial China.* Hong Kong: Kelly & Walsh.

Li, H. 1999: Language test development and College English Test – Spoken English Test. *Foreign Languages and Their Teaching* 126: 28–30.

Li, J. 2002: The current College English Test in China: Problems and thoughts. *Foreign Language Education* 23(5): 33–38.

Li, X. and **Wang, L.** 2000: Testing oral English on a mass scale: Is it feasible? – The oral component of the MET in China. In Berry, V. and Lewkowicz, J., editors, Assessment in Chinese Contexts. Special Issue of the *Hong Kong Journals of Applied Linguistics* 5(1): 160–86.

Lu, D. 2005: A study of the correlation between the extra curriculum reading and the score of Band 6. *Teaching English in China* 28(2): 32–34.

National College English Syllabus for Non-English Majors. 1999: Shanghai, China: Shanghai Foreign Language Education Press.

O'Loughlin, K. 2001: *The equivalence of direct and semi-direct speaking tests. Studies in Language Testing 13*. Cambridge: Cambridge University Press.

Pan, J. 1998: From CET Band Four to college English teaching. *Journal of Tainjin University of Commerce* 2: 63–68.

Qi, L. 2003: *The intended washback of the National Matriculation English Test in China: Intentions and reality.* Unpublished PhD thesis. Hong Kong: The City University of Hong Kong.

—— 2004: Has a high-stakes test produced the intended changes? In Cheng, L., Watanabe, Y and Curtis, A., editors, *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum, 3–17.

—— 2005: Stakeholders' conflicting aims undermine the washback functions of a high-stakes test. *Language Testing* 22(2): 142–73.

Shohamy, E. 1994: The validity of direct versus semi-direct oral tests. *Language Testing* 11(2): 99–123.

Song, X. and **Cheng, L.** 2006: Language learner strategy use and test performance of Chinese learners of English. *Language Assessment Quarterly: An International Journal* 3(3): 241–66.

Sun, C. 2000: Modern language testing and test analysis. *Journal of PLA University of Foreign Languages* 23(4): 82–86.

Wang, H. and **Zhou, F.** 2005: A validation of CET for testing candidates communicative competence and a proposal of a university-based communicative test. *Teaching English in China* 28(2): 26–31.

Wang, Z. 1996: An empirical research on correct–incorrect items in language testing. *Foreign Language Teaching and Research* 28(2): 54–60.

Xiao, Y. 1999: Quality education and College English Test innovations. *Foreign Language Education* 82(4): 72–77.

Xiao, Y. and **Luo, X.** 2002: College English Test innovations. *Foreign Language Teaching and Research* 34(4): 294–98.

Xiong, D., Chen, Y., Liu, Z. and **Huang, G.** 2002: A study of large-scale college English recording oral test. *Foreign Language Teaching and Research* 34(4): 283–87.

Yang, H. 1999: Principles for test design of the College English Test – Spoken English Test (CET–SET). *Foreign Language World* 75(3): 48–57.

Yang, H. and **Jin, Y.** 2000: Score interpretation of CET. *Proceedings at the Third International Conference on English Language Testing in Asia.* Hong Kong: Hong Kong Examinations Authority, 32–40.

Yang, H. and **Weir, C.** 2001: *Validation study of the National College English Test*, third edition. Shanghai, China: Shanghai Foreign Language Education Press.

Yang, X. 2005: "Dumb English" and the reformation of college English teaching. *Journal of Zhengzhou Institute of Aeronautical Industry Management (Social Science Edition)*, 6.

Yu, J. 2005: Assessment and fuzzy comprehension grading of students' English competence. *Teaching English in China* 28(3): 70–75.

Zeng, Y. 2002a: Self-confidence and language test performance. *Modern Foreign Languages* 25(2): 204–09.

—— 2002b: A preliminary study on individualized self-adaptive testing. *Foreign Language Teaching and Research* 43(4): 278–320.

Zhang, H. 2004: *Strategies, approaches to learning, and language proficiency as predictors of EFL reading comprehension*. Unpublished Master's of Education thesis. Kingston, Ontario: Queen's University.

Zhang, H.C. 1988: Psychological measurement in China. *International Journal of Psychology* 23, 101–17.

Zhang, X. 2003: 论大学英语四级考试的负面影响 [On the negative impact of College English Test Band 4], *Educational Science* 19(1): 35–39.

—— 2004: Questioning College English Test. *Foreign Language World* 100(2): 65–69.

Zhao, J. 2006: *Exploring the relationship between Chinese Students' attitudes toward College English Test and their test performance*. Unpublished Master of Education thesis. Queen's University, Kingston, Ontario, Canada.

Zhou, Y. 2004: Comparability study of two National EFL Tests (CET-6 and TEM-4) in China. *The Journal of Asia TEFL* 1(1): 75–100.

Zou, S. 2003: The alignment of language teaching syllabus and language testing: TEM8 test development and administration. *Foreign Language World* 98(6): 71–78.