Pan-Pacific Association of Applied Linguistics 15(2), 71-83.

Fundamental Concerns in High-Stakes Language Testing: The Case of the College English Test

Yan Jin

Shanghai Jiao Tong University

Jin, Y. (2011). Fundamental concerns in high-stakes language testing: The case of the College English Test. *Journal of Pan-Pacific Association of Applied Linguistics*, *15*(2), 71-83.

The College English Test (CET) is an English language test designed for educational purposes, administered on a very large scale, and used for making high-stakes decisions. This paper discusses the key issues facing the CET during the course of its development in the past two decades. It argues that the most fundamental and critical concerns of large-scale high-stakes testing are test validity and fairness as defined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). The CET has a current annual test population of over 18 million, and the results of the test, intentionally or unintentionally, may affect university graduates' employment opportunities, the conferment of a bachelor's degree, and the granting of a residence permit in big cities. The CET test developer, therefore, has been taking measures to make sure that no test taker will be potentially disadvantaged by such factors as test content, test condition, response mode and format, scoring of constructed-response items, and use of test results. Considerable care has been given to the test's validity as well as its operational standardization, which is critical to fairness in high-stakes testing.

The paper begins with an overview of the major developmental stages of the CET since its inception in 1987 and the standardized procedures involved in the CET design, item construction, test administration, test form equation, scoring and score reporting. Following the introductory part, the paper discusses in turn the CET validation efforts in the late 1990s, major revisions of the test with a view to aligning its content and task format with the College English curriculum requirements, and the recent research on the validity of the newly developed internet-based CET, a central focus of which has been on possible biases against test takers who are less proficient in computer operation. Validity and fairness, however, cannot be exclusively addressed in psychometric and technical terms. The use of the test in a particular social context or with particular groups of test takers may be valid and fair or invalid and unfair. In the final part, the paper concludes with a brief discussion of the political dimension of high-stakes testing, with a special

focus on Messick's (1992) unified construct validity argument, which views validity not as a feature or a possession of a test, but a process to validate in a multifaceted approach the uses and interpretations of tests and their scores (Davies, 2003).

Key Words: high-stakes language testing, College English Test, validity, fairness

1 Introduction

Validity and fairness of language tests and testing practices have always been a central concern among language test developers and test users. The 19th Language Testing Research Colloquium, the annual conference of the International Language Testing Association, had 'Fairness in Language Testing' as its theme (Kunnan, 2000). *Language Testing* (1997/14/3) and *Language Assessment Quarterly* (2004/1/2&3), the two scholarly journals in the field of language testing and assessment, dedicated two special issues to the discussion of ethics and professional standards in language testing; *Language Testing* (2010/27/2) recently commissioned several articles debating the conceptualization and frameworks of test fairness and the fairness-validity relation. In this paper, I will use the College English Test (CET), a test of English as a Foreign Language (EFL) in existence for 24 years in China, as an example to illustrate that the most fundamental and critical concerns of large-scale high-stakes testing are test validity and fairness as defined in the *Standards for Educational and Psychological Testing* (hereafter *the 1999 Standards*, AERA, APA, & NCME, 1999).

2 An Overview of the CET: Growing Impact and Increasingly High Stakes

The CET was designed as an end-of-course exit test for non-English-major students in tertiary institutions in China (see CET Design Group, 1987; 1989; 1994a; 1994b; National College English Testing Committee, 2006a; 2006b). During the first two years of their undergraduate study, non-English-major college students are required to take the College English course as part of their curriculum requirements. The course was started in the mid-1980s as a response to the social need for college graduates proficient in English. In the late 1980s, the National College English Testing Committee (NCETC, CET Design Group before 1994) launched the CET Band 4 (CET-4) and Band 6 (CET-6). The testing program has been implemented nationwide since its inception, functioning mainly as a measure to assess the English proficiency level of EFL learners in tertiary institutions in China. The other explicit rationale for the testing program was to promote the implementation of the College English Teaching Syllabus (see Working Group on College English

Teaching Syllabus, 1985; 1986; 1999) and the subsequent College English Curriculum Requirements (Higher Education Department of the Ministry of Education, 2007).

In the past two decades, the CET has gone through several stages of development and major revisions. From the late 1980s till the mid-1990s, the CET Design Group established the standardized procedures involved in the CET design, item construction, test administration, test form equation, scoring and score reporting. From 1993 to 1996, the CET Design Group (after 1994 the NCETC), in collaboration with the British Council, conducted a comprehensive study to validate the test in terms of its content validity, stakeholders' perceptions of the test, concurrent validity with College English teachers' ranking and other external criteria (see Yang & Weir, 1998). The three-year validation efforts identified some weak links in the test's design and the consequential impact on teaching and learning, which led to a series of important decisions to revise the test's format and content and to start the CET Spoken English Test (CET-SET) in the late 1990s. Upon entering the new century, further revisions were made to the test's design and score reporting scheme with a view to better aligning its content and task format with the newly implemented College English Curriculum Requirements (Higher Education Department of the Ministry of Education, 2007). Since 2008, the NCETC has been focusing on the development of the internet-based CET (IB CET) and validation of its construct validity.

All these efforts over the years have steadily improved the measurement quality of the CET and the test has won social recognition among the stakeholders (Jin & Yang, 2006; Yang, 2003). An evidence of its growing popularity was that the past two decades has witnessed a sharp increase in the test population, soaring from some 100,000 in 1987 when CET-4 was inaugurated to the current over 18 million annually. Meanwhile, the results of the test, intentionally or unintentionally, are being used for making increasingly important decisions such as college graduates' employment opportunities, the conferment of a bachelor's degree, and the granting of a residence permit in major cities in China. According to Kane (2002), the stakes of a test come from the consequences of using the test score to make decisions. When these decisions have potentially serious consequences, the testing program is said to involve 'high stakes'. The value accorded to the CET has vastly increased the stakes of the test, which was originally intended to be an optional test for low-stakes educational purposes (see Jin, 2008). This large-scale high-stakes EFL test in China has since attracted growing attention from learners and teachers of College English, educational administrators at different levels, users of various sectors such as employers and government policy-makers (see Jin, 2010; Zheng & Cheng, 2008).

In the next part, I will first discuss the definition of test validity and fairness and the relation between the two and then provide a review of the measures taken by the NCETC to meet the challenges concerning the validity

and fairness of a high-stakes language testing program.

3 Reforms and Revisions: Meeting the Challenges Facing the CET

3.1 Validity and fairness

The 1999 Standards (AERA, APA, & NCME, 1999, p. 9) defines test validity as 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.' The *Standards* conceptualizes fairness as a test quality directly linked to test validity, covering the following four major aspects, equitable treatment of all examinees, freedom from bias, equality of testing outcomes, and equity of opportunity to learn the testing content (ibid.: pp.73-74).

In the language testing community, there in fact have been vigorous debates on the scope of test fairness and its relation to test validity, with some seeing fairness and validity as separate and fairness as an independent test quality, some arguing for fairness as an all-encompassing test quality which subsumes and goes beyond validity, that is, a test cannot be fair if it is not valid, and some stressing that fairness is subordinate to validity, that is, a test has to be fair to be valid (Davies, 2010; Kunnan, 2010; Xi, 2010). Kane (2010, p.177) rightly pointed out that 'the relationship between validity and fairness depends on how we define these two concepts, and perhaps more to the point how broadly we define each of these concepts.' In Kane's view, validity and fairness are intertwined and can be seen as the same question from somewhat different perspectives and involving different emphases, but the overlap between the two is more pronounced than the differences. Kane (ibid.: pp.178-179) made a further distinction between 'procedural fairness', which corresponds to the first two aspects of fairness defined in *the 1999 Standards*, and 'substantive fairness', which includes all of the issues subsumed under *the 1999 Standards'* third and fourth aspects of fairness.

The perspective Kane (2010) adopted with regard to the relation between validity and fairness is intrinsically in line with the way *the 1999 Standards* handles the two most essential aspects of language testing, which, in my view, provides language test developers and test users with useful and practical guidance on how to maintain scientifically justifiable standards of a language test, make meaningful interpretations of test scores, and promote ethically appropriate uses of the test. In the following sections, I will provide a review of the measures taken by the NCETC to address the challenges facing the CET with focuses on the revisions of the test content and format, the standardization of the operational procedures, and the development of the CET Spoken English Test and the internet-based CET.

3.2 Revisions of test content and format

Guided by the development of linguistic theories and in accordance with the requirements set out in the previous College English Teaching Syllabus and the current College English Curriculum Requirements, the NCETC has been continuously revising the content and format of the CET in pursuit of higher validity and more positive impact on teaching and learning. The early version of the CET, for example, relied heavily on the objective format of multiple choice questions (MCQ) and had a significant proportion of discrete-point items assessing knowledge about language such as grammar and vocabulary. The CET validation study in the mid-1990s pinned down the limitations of over-reliance on the MCQ format and discrete-point tasks and introduced a number of constructed-response item types to the CET in the late 1990s, including, for example, compound dictation (for words and sentence chunks), short answer questions, translation (from English to Chinese and from Chinese to English). In the latest 2007 revision of the CET, the proportions of the test's component parts were adjusted and a new section of fast reading, or skimming and scanning (with strict time control), was added.

3.3 Standardization of operational procedures

Since its inception in 1987, the CET has established and followed standardized operational procedures in every aspect of the testing process: item construction, test administration, test form equation, scoring and score reporting. Take item construction as an example. Though one test paper is officially released every two years, all the test items are actually exposed right after being used in live tests. Therefore, new items have to be written for each test administration. A series of measures have been taken and standardized procedures followed to ensure the quality of the items, which include: 1) regular training of item writers; 2) item review groups reviewing the items submitted by the item writers for the appropriateness of source materials and the quality of the items; 3) pilot-testing each item (except the writing task) among a representative group of prospective test takers; 4) analyzing item statistics to check the quality of items and detect potential biases against test takers with different backgrounds; 5) further modifications to accepted items by expert item reviewers; 6) construction of test papers based on test specifications; 7) native speakers of English reviewing the draft version of the test papers to perfect the language; 8) the NCETC members reviewing the final version of the test papers. Typically, it takes over a year for a raw item to undergo these stages of careful scrutiny and be accepted, pilot-tested, revised and finally used in the live test. Post-hoc data analysis will also inform item writers of the performance of each item with a representative group of test takers.

Test administration is organized through a hierarchical operational

structure with the participating institutions (colleges and universities) playing a major role in student registration and test invigilation, the provincial/municipal examinations authorities supervising the institutions, and the National Educational Examinations Authority taking charge of the entire operation. Accommodation is provided at the request of test takers with special requirements. Test takers who are visually impaired, for example, are accommodated with an enlarged version of test papers, and the hearing-impaired with lip reading for the listening test.

As for scoring, systematic procedures of quality control have been established and effective measures taken to ensure the inter- and intra-rater reliabilities and the inter-center reliability of constructed-response items, including dictation, short answer questions, translation, error correction, sentence completion, and essay writing. Since the CET online marking system was put into operation in 2006, both the efficiency and quality of marking have been greatly improved. In terms of score reporting, a norm-referenced approach is adopted for the CET score interpretation (see Yang & Jin, 2001). To make sure that the scores of different test forms are comparable, the NCETC uses anchored test takers to adjust the difficulty level of every form of the test. That is, the difficulty level of a new form of the test is equated to that of an equation test by means of linking performances of a representative sample of test takers who take both the live test and the equation test. Therefore, scores of each administration are equated and normalized before being reported to test takers. A variety of descriptive data are also provided to participating institutions and educational authorities for educational evaluation purposes (see Jin, 2011a).

3.4 The CET Spoken English Test

The project of developing the CET Spoken English Test (CET-SET) was initiated in the mid-1990s, when the NCETC became aware of an increasing need for college graduates with a better ability to speak English as a foreign language. The CET-SET was officially launched in 1999 in four major cities in China with a test population of several hundred. Taking a face-to-face interview format with two examiners and three candidates forming a test group, the CET-SET engages test takers in a number of monologic and interactive tasks such as question-and-answer, individual presentation and group discussion. An analytic approach is adopted for the scoring of the following three aspects of candidates' performances in the test: 1) the quality of the language (accuracy and range), 2) contribution to the interaction and the cohesion and coherence of the discourse (size and discourse management), and 3) flexibility to deal with different topics and use of communicative strategies (flexibility and appropriacy). The weighted total of the three sub-scores is converted into a grade and reported to the test taker (National College English Testing Committee, 1999).

Teachers' and students' perceptions of the CET-SET were on the whole very positive and beneficial effects on teaching and learning were envisaged by a noticeable shift of attention to teaching speaking in College English classes (Jin, 2000). By 2010, a total of 58 CET-SET test centers have been established, which can accommodate a maximum of 100,000 test takers a year. The test, however, is only accessible to students who have achieved a CET-4 score of 550 or above, or a CET-6 score of 520 or above due simply to the constraints imposed by the labor-intensive format. The NCETC, therefore, has designed a computer-based version of the CET-SET, which is capable of assigning test takers to random groups in a test room and engaging group members in interactive discussions. The computer-based CET-SET was implemented on a trial basis in May 2011.

3.5 The internet-based CET

In an era of extensive application of the internet and computers, target language use (TLU) situations have changed in fundamental ways. Computer-mediated communication has become a major characteristic of TLU situations. Accordingly, language tests should make full use of technological innovations to incorporate the major features of computer-mediated communication into the design of test tasks and user interfaces. The project of the internet-based CET (IB CET) was initiated against such a background, in addition to the practical needs for improving the efficiency of test administration and looking for a solution to the thorny issue of high-tech cheating which poses a direct threat to the validity of the paper-based CET.

The trial implementation of the IB CET-4 took place in December 2008 and the IB CET-6 in December 2009. Distinctive features of the design of the IB CET are as follows: 1) tasks of an integrated nature are employed, which engage test takers in multi-modality language activities; 2) speaking is included as a new component; 3) authentic audio-visual clips are used as listening test materials; 4) all the tasks are completed on the computer, which requires test takers to read on the screen, view video clips, listen to audio clips, record their voices, click and double click the mouse to select or de-select an answer, drag and drop an option, and type out answers for dictation, sentence completion, and essay writing.

A major fairness concern with this innovative way of testing is that test performances on the IB CET might be affected by the level of test takers' computer proficiency. Studies were therefore conducted to investigate the effect of test takers' computer familiarity and anxiety on test performance, and the effect of test modes (paper- versus the internet-based) on the processes involved in essay writing and the texts produced. These studies identified a statistically significant relationship between test takers' computer familiarity and anxiety and test performance. The statistics of the effect size confirmed a practical significance of the effect of computer familiarity and anxiety on test

performance, which seems to indicate that test takers less proficient in computer operation could be disadvantaged by the new way of testing. The analysis of the texts produced in the paper- and internet-based tests, however, showed that test takers could produce significantly lengthier and syntactically more complex texts when writing on the computer. The analysis of the writing process also confirmed that with the increase of computer familiarity, the use of cognitive strategies involved in essay writing also improves.

It was argued that in a language test in the 21st century, computer literacy should no longer be considered as a source of construct-irrelevant variance; instead, it has become an important type of test taker attribute that interacts with test task characteristics (Jin, 2011b). The interaction may enhance or impede test takers' performances of computer-mediated language tasks. For a better understanding of the nature of interaction in the construct of a computer- or internet-based language test, Chalhoub-Deville's (2003) local, context-bound view of language ability is considered relevant. Unlike the conceptualization of a global construct, which views interaction in language use from an individual-focused cognitive perspective, the stance taken by Bachman (1990) in his well-known CLA model, a local construct adopts a social interactional perspective, that is, individual ability and contextual facets interact in ways that change them both. This social-cognitive construct representation is useful for a better understanding of the IB CET construct. Quoting Brown's (2003) study of interviewer variation and the construct of a speaking test, Chalhoub-Deville (2003, p.378) stressed that 'it is, it seems, simply not appropriate to assume that the variation that is allowed to occur is not relevant to the construct… I would even argue that variation is inevitable if we view ability within context as the construct.' For future studies of the technology-enhanced way of language testing, a research agenda should therefore be set out for a clearer definition of computer literacy for language use and better ways of engaging test takers' computer literacy to facilitate test performance.

3.6 A summary

The table below summarizes the efforts made by the CET Design Group and the NCETC in the past two decades to address the changing social needs and target language use situations at the different stages of the CET development, most of which have been discussed in the brief review above.

Table 1. Challenges Facing the CET and Measures Taken to Cope with the Challenges

Timeframe	Social needs and TLU situations	Things done and measures taken
CET-4 and CET-6		

Mid-1980s ~Late 1980s	[1]	Social needs for university graduates proficient in English	[1]	Needs analysis, design and development of the CET
	[2]	Promulgation of College English Teaching Syllabus in 1985/1986	[2]	Launch of the CET-4 in 1987 and the CET-6 in 1989
Early 1990s	[3]	Score comparability and interpretability	[3]	Test form equating and establishing the CET norm
	[4]	Procedural standardization	[4]	Establishing CET operational procedures
Mid-1990s ~Late 1990s	[5]	Increasing recognition of the CET by stakeholders and growing impact on College English teaching and learning	[5]	CET validation study; introducing new item types to the CET and reporting Grade Point Average to institutions
Early 2000s ~Mid-2000s	[6]	Increasingly higher stakes of the CET resulting in teaching to the test and over- or misuses of the CET	[6]	New score reporting scheme since 2005 and major revisions to the design of the CET and its content and format
Late 2000s ~now	[7]	Ethical concerns with high-stakes testing and professionalism in the language testing community	[7]	Survey of EFL testing practices and developing and validating Code of Practice for EFL tests in China
CET-SET				
Mid-1990s ~Late 1990s	[1]	Social needs for higher proficiency in spoken English	[1]	Needs analysis, design and development of the CET-SET
	[2]	Promulgation of revised College English Teaching Syllabus in 1999	[2]	Launch of the CET-SET in 1999
Early 2000s ~Mid-2000s	[3]	Procedural standardization and quality control of marking	[3]	Establishing CET-SET operational procedures including examiner training
	[4]	Increasing number of test takers	[4]	Setting up CET-SET test centers
Late 2000s ~now	[5]	Extensive use of the internet and computers in academic and social life	[5]	Trial implementation of computer-based CET-SET in 2011
IB CET-4 and IB CET-6				
Late 2000s ~now	[1]	Extensive use of the internet and computers in academic and social life	[1]	Needs analysis and design of the IB CET and user interfaces
	[2]	Promulgation of College English Curriculum Requirements in 2007	[2]	Trial implementation of IB CET-4 in 2008 and IB CET-6 in 2009
	[3]	High-tech cheating in paper-based test	[3]	Developing the IB CET item bank
	[4]	Effects of computer proficiency on test	[4]	Empirical investigation of the effects and theoretical

performance	conceptualization of the construct being measured

4 The Way Forward: Working towards a Code of Practice

In his discussion of the three heresies of language testing research, Davies (2003, p.363) supported Messick's (1992) unified construct validity argument which views validity not as a feature or a possession of a test but a process to validate in a multifaceted approach the uses and interpretations of tests and their scores. Quoting Cronbach (1971) and Messick (1989), Kane (2002, p. 31) added a similar annotation to the definition of test validity provided in *the 1999 Standards*: 'The test itself is not validated, and test scores per se are not validated. It is the interpretation determined by the proposed use that is validated.' Kane (ibid.: p. 32) made a useful distinction between 'descriptive interpretations' and 'decision-based interpretations', and pointed out that 'the proponents of the testing program focus their attention on a content-based interpretation..., while taking the appropriateness of the test use for granted. On the other hand, the critics often focus on the consequences of testing programs and on the value judgments implicit in the decisions being made.'

Validity and fairness, therefore, cannot be exclusively addressed in psychometric and technical terms. The use of the test in a particular social context or with particular groups of test takers may be valid and fair or invalid and unfair. As pointed out by Davies (2003, p.361), 'Tests are inevitably political since what they do – in education as in immigration – is to sort and select to meet society's purposes. Testers cannot expect that their work will not have a political dimension. The proper reaction to such concern is surely to act with professional skill and rectitude within the contexts in which they work.'

As part of the research project sponsored by the Education Commission of the Shanghai Municipal Government to develop a code of practice for EFL test developers and users, a survey of large-scale high-stakes EFL tests and testing practices in China was recently conducted with respect to test development, administration and use (Fan & Jin, 2010; 2011). Synthesizing the views from test developers, including representatives from six predominant EFL examination boards in China, and the primary stakeholders of these tests, including 166 EFL teachers and 490 students from different regions of the country, the study reached the conclusion that examination boards on the whole follow their own quality control procedures in developing, administering and validating their tests. But the validity of these procedures is open to question. Over-uses or misuses of EFL tests were identified as having constituted a serious threat to test validity.

The study awakens China's language testers to the importance and urgency of developing a code of practice which is applicable to China's EFL

testing context and also calls for more communication between test developers and stakeholders. Test validity and fairness, the most fundamental concerns in high-stakes language testing, are the joint responsibility of all stakeholders in the testing process. Though it is premature to prescribe enforcement mechanisms in such a code of practice, the purposes of the code, at the present stage, are mainly educational and inspirational, or to be specific, to raise the awareness of professionalism and quality among the EFL test developers in China, and to communicate to the stakeholder groups the basics of language testing and good testing practices.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1-25.

CET Design Group. (1987). *College English Test Band 4 Syllabus and Sample Tests*. Shanghai: Shanghai Foreign Language Education Press.

_____. (1989). *College English Test Band 6 Syllabus and Sample Tests*. Shanghai: Shanghai Foreign Language Education Press.

_____. (1994a). *College English test Band Four Syllabus and Sample Tests (Revised edition)*. Shanghai: Shanghai Foreign Language Education Press.

_____. (1994b). *College English Test Band Six Syllabus and Sample Tests (Revised edition)*. Shanghai: Shanghai Foreign Language Education Press.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing*, *20*(4), 369-383.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (ed.), *Educational measurement (2nd ed.)* (pp. 443-507), Washington, D.C.: American Council on Education.

Davies, A. (2003). Three heresies of language testing research. *Language Testing*, *20*(4), 355-368.

_____. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171-176.

Fan, J., & Y. Jin. (2010). Standards for language testing: review, reflection and implications. *Foreign Language World*, 1, 82-91.

_____. (2011). *The way towards a code of practice: A survey of EFL*

testing in China. Paper presented at the 33rd Language Testing Research Colloquium held in Ann Arbor, Michigan, June 23-25, 2011.

Higher Education Department of the Ministry of Education. (2007). *College English Curriculum Requirements*. Shanghai: Shanghai Foreign Language Education Press.

Jin, Y. (2000). Washback effects of CET-SET on EFL teaching in China. *Foreign Language World*, 4, 56-61.

_____. (2008). Powerful tests, powerless test designers?—Challenges facing the College English Test. *English Language Teaching in China*, *31*(5), 3-11.

_____. (2010). The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.). *English Language Assessment and the Chinese Learner*. (pp. 44-59). Routledge, Taylor & Francis Group.

_____. (2011a). On the educational evaluative value of standardized language tests. *Foreign Language Testing and Teaching*, *1*, 26-34, 64.

_____. (2011b). *Is computer literacy construct-relevant in a language test in the 21st century?* Paper presented at the 33rd Language Testing Research Colloquium held in Ann Arbor, Michigan, June 23-25, 2011.

Jin, Y., & H. Yang. (2006). The English Proficiency of College and University Students in China: As Reflected in the CET. *Language, Culture and Curriculum*, *19*(1), 21-36.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

_____. (2010). Validity and fairness. *Language Testing*, *27*(2), 177-182.

Kunnan, A. (Ed.). (2000). *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Kunnan, A. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*(2), 183-189.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement (3rd ed.)*, (pp. 13-103). New York: American Council on Education and Macmillan.

_____. (1992). Validity of test interpretation and use. In Alkin, M. C. (Ed.), *Encyclopedia of educational research*. (pp. 1487-1495). Sixth edition. New York: Macmillan.

National College English Testing Committee. (1999). *CET-SET Test Syllabus and Sample Tests.* Shanghai: Shanghai Foreign Language Education Press.

_____. (2006a). *CET-4 Test Syllabus and Sample Tests* (Revised edition). Shanghai: Shanghai Foreign Language Education Press.

_____. (2006b). *CET-6 Test Syllabus and Sample Tests* (Revised edition). Shanghai: Shanghai Foreign Language Education Press.

Working Group on College English Teaching Syllabus. (1985). *College English Teaching Syllabus (Science and Technology)*. Shanghai: Shanghai Foreign Language Education Press.

_____. (1986). *College English Teaching Syllabus (Arts and Science)*. Shanghai: Shanghai Foreign Language Education Press.

_____. (1999). *College English Teaching Syllabus (Revised edition)*. Shanghai: Shanghai Foreign Language Education Press.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147-170.

Yang, H. (2003). A review of the development of the College English Test in the past fifteen years. *Journal of Foreign Languages*, *3*, 21-29.

Yang, H., & C. Weir. (1998). *CET Validation*. Shanghai: Shanghai Foreign Language Education Press.

Yang, H., & Y. Jin. (2001). CET score interpretation. *Foreign Language World*, *1*, 62-68.

Zheng, Y., & L. Cheng. (2008). College English Test (CET) in China. *Language Testing*, 25 (3), 408-417.

Yan Jin
Shanghai Jiao Tong University
Haoran Hi-tech Building 2203
Phone：86 21 62812890
Fax: 86 21 62932756
E-mail: yjin@sjtu.edu.cn

Received: August 4, 2011
Revised: November 17, 2011
Accepted: December 7, 2011