



Yu, G., & Zhang, J. (2017). Computer-Based English Language Testing in China: Present and Future. *Language Assessment Quarterly*, 14(2), 177-188.
<https://doi.org/10.1080/15434303.2017.1303704>

Peer reviewed version

Link to published version (if available):
[10.1080/15434303.2017.1303704](https://doi.org/10.1080/15434303.2017.1303704)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Taylor & Francis at <http://www.tandfonline.com/doi/abs/10.1080/15434303.2017.1303704>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Computer-based English Language Testing in China: Present and Future

Abstract

In this special issue on high-stakes English language testing in China, the two articles on computer-based testing (Jin & Yan; He & Min) highlight a number of consistent, ongoing challenges and concerns in the development and implementation of the nation-wide IB-CET (Internet Based College English Test) and institutional computer-adaptive English tests, respectively: conceptualizing the construct of computer-based language testing, ensuring fairness for test takers with differing levels of computer literacy, and achieving comparability between tests or tasks of different delivery modes. In this article, we provide an overview of the research studies on computer-based English language testing conducted by Chinese scholars and published in major Chinese academic journals in recent decades, aiming to identify the research topics, gaps, and agendas that could have implications beyond Chinese contexts in promoting better use of computer technologies *in* and *for* English language testing.

Key words: Computer, China, language testing research

Special issues of international journals like *Language Assessment Quarterly* that focus on the regional development (e.g., Taiwan, Volume 9, Issue 1, 2012; mainland China, this issue; Japan, forthcoming) of English language testing are long overdue, given the increasing globalization and localization of high-stakes English language tests. Large-scale English language tests in China currently include international tests such as the Test of English as a Foreign Language, Internet-based Test (TOEFL iBT) and International English Language Testing System (IELTS), and numerous nationally or locally developed tests such as the College English Test (CET), Test for English Majors (TEM), Graduate School Entrance English Examination (GSEEE), Medical English Test System (METS) for nurses, National Matriculation English Test (NMET) for secondary school graduates, and Public English Testing System (PETS) for the general public (see Cheng, 2008 for brief introductions to the locally developed tests except for the METS). Among these locally developed tests, the CET is now Internet-based, and the speaking components of the PETS (Level 1, from September 2006) and NMET (in some provinces, e.g., Guangdong and Guangxi) are computer-based (Zeng, 2010). In addition, a few universities offer computer-based English tests to their students, often using commercially available platforms. Certain questions emerge in view of these developments: What is the current status of research on computer-based English language testing in China? To what extent are the current practices of computer-based English language testing in China supported by sufficient and strong research evidence from its unique social and educational contexts?

1
2
3 A systematic search that we conducted on the China Knowledge Resource Integrated
4 Database¹--using terms (all in Chinese) such as “computer/Internet-based”, “computer-
5 adaptive”, “computer-delivered”, “computer-aided English/language test”, and “automatic
6 scoring”—revealed that there are a good number of articles on computer-based English
7 language testing published in Chinese academic journals. The vast majority of them,
8 however, were think-pieces introducing, reviewing, or debating the challenges and
9 potentials of computer-based English language tests and automated scoring systems; only a
10 small number of publications reported empirical studies. Although these empirical studies
11 are not typically of the depth or quality of Jin and Yan’s or He and Min’s (in this issue), this
12 brief review of these studies sheds light on the current status of research on computer-
13 based English tests in China. More importantly, the current review identifies research gaps
14 and points to future research agendas, which can have implications beyond Chinese contexts
15 to promote better use of computer technologies *in* and *for* language testing.
16
17
18
19
20

21 The Present

22
23
24 Three important research topics have appeared in publications on computer-based English
25 language testing in China. The first has focused on computer-adaptive testing, driven to a
26 great extent by the promising efficiency of test delivery, but constantly challenged by
27 difficulties in designing appropriate techniques to assign the optimal number of items of
28 appropriate difficulty levels to test takers. The second body of research, and arguably the
29 largest in terms of number of publications, is related to the College English Test and
30 achievement tests designed and used in individual universities via commercially available
31 test systems. The third is concerned with the development of automated scoring systems to
32 assess speaking, writing, and translation.
33
34
35

36
37 In terms of publication venues, *Foreign Language World* [外语界], *Computer-Assisted*
38 *Foreign Language Education* [外语电化教学], and *Foreign Language Testing and Teaching*
39 *[外语测试与教学]* (all published by Shanghai International Studies University) have been
40 the three major journals publishing these empirical studies, followed by *Modern Foreign*
41 *Language* [现代外语] (published by Guangdong University of Foreign Studies) and *Foreign*
42 *Language Teaching and Research* [外语教学与研究] (published by Beijing Foreign Studies
43 University). Other journals, such as the *China Information Processing Journal* [中文信息学报]
44 and the *Journal of Tsinghua University* [清华大学学报: 自然科学版], have mainly published
45 studies on automated scoring systems.
46
47
48
49
50

51 Research on Computer-adaptive Testing

52
53
54 The earliest attempts to develop and research computer-based English tests in China started
55 in Guangdong University of Foreign Studies in the 1990s and focused mainly on computer-
56 adaptive testing. A number of articles were published by staff or former students of this
57
58
59
60

1
2
3 institution, for example, He (1999), Zhang (1999), Zeng (2002), and Huang and He (2013). He
4 (1999) compared 55 university students' performance in a conventional paper-based test to
5 their teachers' rankings of their language ability and to their performance on a "cognitive
6 computer-adaptive test", which included reading comprehension, grammar and vocabulary,
7 and cloze items. He (1999) found that the "cognitive computer-adaptive test" was more
8 efficient, accurate, and consistent in assessing the participants' language abilities than was
9 the conventional paper-based test. Zhang (1999) reported a high correlation ($r=.86$) between
10 a computer-adaptive and a "self-adaptive" test which allowed the participants to decide the
11 difficulty level of the next item presented to them. In Zhang's project, 50 university students
12 completed multiple-choice vocabulary items. In addition to computer-adaptive and "self-
13 adaptive," Zeng (2002) piloted "individualized self-adaptive testing" which required test
14 takers not only to decide the difficulty level of the next item but also to report their
15 confidence level in completing the test items (incorporated as a weighting element in
16 scoring the participants' test performance). Zeng (2002) also reported that the addition of
17 self-assessed confidence scores had some advantage over the use of maximum likelihood
18 estimation in computer-adaptive testing.
19
20
21
22
23

24 Recently, Huang and He (2013) conducted Monte Carlo simulations to assess the usefulness
25 of a three-parameter logistic graded model to address issues of local item dependence in a
26 testlet in a computer-adaptive test of listening comprehension. Instead of statistical
27 simulations, He and Min (this issue) used students' actual performance data in a
28 conventional computer-based test and a corresponding computer-adaptive test, which
29 included both dichotomously-scored items and polytomously-scored testlet-based items of
30 reading and listening comprehension. Over 8,200 students took the conventional computer-
31 based test, with 416 of them taking the computer-adaptive test, and completing a
32 questionnaire designed to collect data such as the students' background profiles, their
33 paper-based language test scores (in particular, on the CET), and indicators of computer
34 familiarity. Shortly after taking the computer-adaptive test, some students voluntarily
35 participated in focus-group discussions. He and Min reported that the conventional
36 computer-based test and the computer-adaptive test were comparable and measured the
37 same construct. Moreover, students' English language proficiency as measured by the
38 paper-based test, unlike their computer familiarity, was a significant predictor of their
39 performance in the two computer-based tests. Furthermore, factorial invariance of the
40 computer-adaptive test scores of male and female students was noted, as another piece of
41 supporting evidence of the quality of this computer-adaptive test, which is taken by
42 undergraduates from Zhejiang University as part of their graduation requirement.
43
44
45
46
47

48 **Research on the Computer-based CET and Institutional Achievement** 49 **Tests at Universities** 50

51
52
53 Du and Gui (2000) were among the first in China to develop their own system to explore the
54 usefulness of computer-based English language tests alongside others cited immediately
55 above. There were some sporadic efforts to develop computer-based English language testsⁱⁱ
56 in China between 2000 and 2005. However, it was the announcement in February 2005 (全
57
58
59
60

1
2
3 国大学英语四、六级考试改革方案<试行>) that the National CET Committee would
4 consider using a computer-based CET that inspired, pushed, or influenced various
5 stakeholders (e.g., teachers, researchers, universities, and publishers) to develop computer-
6 based tests and testing systems. It was the commercial availability of a few computer-based
7 language testing systems in the Chinese market (e.g., College English Oral Test Systemⁱⁱⁱ of
8 Shanghai Foreign Language Education Press, iflytek^{iv} as a spin-out publically listed company
9 of the University of Science and Technology of China, Lange^v of the Lancoo Group, and
10 Wingsoft^{vi} of Fudan University) that made it possible for universities and English language
11 teachers to develop their own computer-based tests to assess their students on a large scale
12 efficiently, because these systems often included a set of components for item development,
13 delivery, marking, and data analysis and reporting.
14
15
16
17

18 Since 2005, a number of studies have reported local attempts to develop and use computer-
19 based language tests in universities. These empirical studies explored a number of validity
20 issues in the institutional computer-based achievement tests and the national IB-CET, such
21 as the comparability between computer-based and paper-based tests in terms of students'
22 performances on tests, the impact of computer literacy on test results and test-taking
23 cognitive processes, the influence of delivery modes on the features of language produced in
24 speaking and writing tasks, students' perceptions of and readiness for computer-based tests,
25 and the advantages and the challenges of using computer-based tests. The majority of these
26 studies have focused on the assessment of speaking and writing abilities, in other words,
27 productive rather than receptive language abilities. This focus presents an interesting
28 contrast with research on computer-adaptive tests which primarily used listening and
29 reading tasks.
30
31
32
33

34 In what follows, we outline the key findings or research focuses of these studies in the
35 chronological order of their publication. Qiu, Ji, Wan and Cheng (2005), using Wingsoft for
36 test delivery at Fudan University, described their students' performances as well as the
37 benefits and challenges in using computerised listening and speaking tasks such as read-
38 aloud, summarization of pictures/cartoons, listening to news from foreign radio stations,
39 and short debates between two test takers. Cai (2005), also using Wingsoft and at Fudan
40 University, compared 182 students' performances in the computer-based and the face-to-
41 face CET-Spoken English Test. He found a reasonable correlation ($r= 0.71$) between the two
42 test modes. Similarly, Gao (2007) reported the comparability of students' performances on
43 two computer-based speaking tests and one face-to-face speaking test at Hangzhou Dianzi
44 University, observing generally positive attitudes of the students towards computer-based
45 speaking tests of English. In a CET sponsored study, Cai and Wang (2009) compared
46 computer-based and paper-based writing tests; they found that participants' typing speed,
47 anxiety in typing, and computer familiarity did not significantly affect their writing
48 performance. They also found that the marks assigned to the computer-processed and the
49 hand-written scripts were highly correlated. From test takers' and teachers' perspectives, Li
50 (2009) (in a study sponsored by the CET) investigated the extent to which the use of
51 different types of speaking tasks (e.g., read-aloud, story completion, describing pictures,
52 listening to retell/summarize, and group discussion) with multimedia input in end-of-term
53 examinations might help to reduce test takers' anxiety differently. The speaking tasks in this
54
55
56
57
58
59
60

1
2
3 study were delivered via the Lange system. The key finding of Li's study was that test takers'
4 anxiety was alleviated by the use of multimedia input – an existing practice of the IB-CET.
5 Zhu and Zhang (2009) reported the benefits and challenges of using computerised listening
6 and speaking tasks as part of the achievement tests at Ningbo University since 2006.
7 Although the delivery platform was developed by their university's education technology
8 company, Zhu and Zhang (2009) modelled their assessment tasks after Qiu et al. (2005) and
9 reported similar findings. Tang and Liu (2009), from Beijing Foreign Studies University,
10 reported that the performances of students with intermediate-level English proficiency were
11 affected by the mode of test delivery, especially for the reading section of their computer-
12 based test. Yin, Zheng, Wang and Xin (2010), from Harbin Institute of Technology, reported
13 the differential effects of two delivery modes on 30 test takers' fluency, operationalized as
14 seven features of speech such as average length of runs, pauses, errors, and subordinate
15 clauses. Three types of tasks were used by Yin et al.: two non-interactive tasks (short-answer
16 questions; answers to questions with supporting information provided) and one interactive
17 task (triadic small group discussions on a given topic). They concluded that the students'
18 fluency in non-interactive tasks was less likely to be affected by test delivery modes, while
19 the interactive task (i.e., small group discussion) was more susceptible to test delivery
20 modes, with significantly more errors and slower speaking rates observed in the computer-
21 based speaking test.
22
23

24
25
26
27 Dai and You (2010) reported the results of Rasch analysis of six raters' severity and
28 consistency in marking over 660 students' performances in three different types of
29 computer-based speaking tasks; they pointed out that rater bias would not go away simply
30 because of the use of computer-based testing. Dai (2011) then reported high comparability
31 between face-to-face oral proficiency interviews (OPI) and computer-based OPI (COPI) and
32 high consistency and similarity between two raters when marking OPI and COPI test
33 performances. However, according to the data from a survey of test takers, the students
34 preferred the OPI and considered COPI tasks less interactive than those in the OPI (see also
35 Qian 2009 which reported similar findings based on data collected from Hong Kong
36 university students). L. Jin (2011) reported on the practice of computer-based speaking tests
37 in Inner Mongolia Normal University since 2005 via the Lange testing system. Xu, Xie, Liu,
38 Chen, Liu and Gu (2013) reported a high correlation ($r=0.91$) between a teacher-
39 administered face-to-face speaking test (involving a short-answer question and a topic-
40 based dialogue between two test takers) and a computer-based speaking test (involving
41 reading aloud a short passage and a topic-based monologue) delivered via iflytek. The
42 students' performances on the computer-based test were automatically assessed through
43 iflytek's automated evaluation system (see below). Based on data from the questionnaire
44 survey given to the students and on interviews with six of their teachers, Xu et al. (2013)
45 reported the generally positive attitudes of these stakeholders towards the computer-based
46 speaking test.
47
48
49
50
51
52

53 The studies reviewed above all focused on English for general purposes. Research on
54 computer-based assessment of English for specific purposes or content-based assessment
55 has been rare in China, except for Si (2008) and Chen (2009). Si (2008) investigated the
56 computer-based assessment of students' speaking ability in business contexts, while Chen
57 (2009) piloted computer-based assessment of students' knowledge of English-Speaking
58
59
60

Countries. He found that the presentation of test materials in multimedia had a negative effect on the validity of the test.

As shown above, there have been several research studies on computer-based English language tests developed by individual universities; however, the nation-wide IB-CET, the largest computer-based English test in China, has not enjoyed the same extent of research effort. The first administration of the IB-CET in 56 universities (with 100 participants maximum from each institution) in December 2008 prompted a few studies designed to investigate the validity of the IB-CET and how well university students were coping with the new test formats. For example, Huang and Qin (2009) surveyed 36 test takers from a southwestern university in the second week after the actual test. They found that the students were generally comfortable with the computer-based test formats, but probably needed more time to adjust themselves to the new listening and speaking tasks because the formats and time pressure of these tasks differed from the paper-based CET. Like Huang and Qin (2009), Liu (2011) surveyed 185 test takers for their views on (a) the key differences between the IB-CET and paper-based tests and (b) their challenges in different sections of the IB-CET. Similar to Huang and Qin (2009), Liu (2011) reported generally positive attitudes towards the IB-CET, but also suggested areas for improvement especially with regard to test content and computer interface. Drawing on data from questionnaire and interviews, Yang and Li (2010) reported a significant negative correlation between 52 test takers' computer anxiety and their perceptions of computer self-efficacy, as well as a significant positive correlation between computer anxiety and test anxiety in computer-based speaking tests. Jin and Yan (this issue) reported that test takers' high computer literacy can facilitate their performance in the IB-CET writing tasks (see also Jin & Wu, 2010) although their cognitive processes involved in the computer-based and paper-based writing tasks were similar. Jin and Yan argued that computer literacy should be considered as a contextual factor "closely related to the construct" being measured in computer-based tests, given the extensive use of computers in everyday life these days.

Research on Automated Scoring: Speaking, Writing and Translation

A limited amount of research in China has focused on three areas of automated scoring and test performance: speaking, writing, and translation. The scope of this article does not permit a discussion of natural language processing, computational linguistics, or statistical linguistics as part of a critique of automated scoring engines. Rather, we focus here on interpreting the reported correlations between the scores generated by the automated engines and the scores assigned by human raters, which is often claimed to be supporting evidence for the quality of automated scoring engines (cf. Chapelle & Douglas, 2006).

To the best of our knowledge, the IB-CET and PETS have implemented automated evaluation systems to assess speaking test performance (for Levels 1 and 2). However, there does not seem to be any published research on the automated evaluation system used in the IB-CET speaking test. For the PETS, Qiao, Dong and Liu (2012) described the components of the

1
2
3 automated scoring engine – EduRater. Based on about 1,000 test takers’ performances in
4 two different PETS-1 speaking tasks, which were marked by three raters independently, Qiao
5 et al. reported a high correlation ($r=0.81$) between the automated and human scoring. Li,
6 Yang, Feng, Wu, Chen and Hu (2008), Yan, Hu, Wei, Dai, Li, Yang and Feng (2009) and Yan,
7 Hu, Wei, Li, Yang and Feng (2010) – a research team from the University of Science and
8 Technology of China – reported on their attempts to develop an automated system^{vii} to
9 evaluate students’ performance on speaking tasks – reading-aloud, retelling/summarization,
10 and recitation, respectively. These three articles reported high correlations between
11 automated and human scoring. However, Zhou and Zeng (2016) reported that automated
12 and human scoring of high-school leavers’ speaking performance differed significantly in
13 terms of rater severity even though the overall distribution of students’ test scores were
14 similar across the two scoring methods. For the automated evaluation of writing
15 performances, several researchers (e.g., Ge, 2010; Ge & Chen, 2009; Li & Ge, 2008; Li & Liu,
16 2013) have proposed slightly different models. Another interesting area of development is
17 the automated evaluation of translations between English and Chinese (Jiang, 2013; Jiang &
18 Wen, 2010, 2012; Liu & Liu, 2015; Wang & Chang, 2009; Wen, Qin & Jiang, 2009). These
19 publications were based on doctoral dissertations, and all reported some kind of “superiority”
20 for their automated scoring engines for evaluating speaking, writing, or translation; however,
21 none of these engines has been externally validated beyond the initial thesis inquiries.
22
23
24
25
26

27
28 These three main research topics on computer-based English language testing demonstrate
29 Chinese researchers’ endeavours to better understand the efficiency of computer-based
30 testing and the comparability between computer-based and paper-based tests alongside
31 their efforts to develop automated scoring engines to evaluate speaking, writing, and
32 translation performance. In addition, researchers have also been concerned about test
33 takers’ readiness for and attitudes towards computer-based tests as well as the fairness of
34 computer-based tests for students of different experience and ability. These studies provide
35 solid stepping stones into the future.
36
37
38

39 The Future

40
41
42
43 Computer technology is being used “in designing, developing and delivering test content as
44 well as scoring and reporting examinee test performance” (Sawaki, 2012, p. 426). Chapelle
45 (2010) listed three motives for using computer technology in language testing: efficiency,
46 equivalence, and innovation. The use of computer technology can improve efficiency in test
47 development, delivery and scoring, for example, via computer-adaptive testing and
48 automated scoring of speaking and writing performances. Equivalence refers to the
49 comparability in test performances between computer-based and paper-based or other
50 traditional methods of assessment. By innovation, Chapelle suggested that the integration of
51 technology can help to reconceptualize language ability as “the ability to select and deploy
52 appropriate language through the technologies that are appropriate for a situation”
53 (Chapelle & Douglas, 2006, p. 107). Douglas (2013, p. 2) urged that “we must define the
54 language construct to include appropriate technology in light of the target situation and test
55 purpose.”
56
57
58
59
60

1
2
3 None of the empirical studies reviewed here has incorporated or viewed technology as part
4 of the construct to be assessed in their tests (but see Jin & Yan, this issue, who call for a
5 reconceptualization of the construct of computer-based writing tests). These past studies
6 have simply been used or considered computer as a tool rather than as an integral aspect of
7 the construct being assessed. Given the extensive use of computer technology in language
8 learning and communication nowadays, especially in higher education and work, it is high
9 time to reconceptualize what is meant by computer-based language testing. In our view,
10 such a reconceptualization should be the premise and the guiding rationale for any
11 innovation in assessment practices. As Chalhoub-Deville (2010, p. 522) contended, "L2 CBTs,
12 as currently conceived, fall short in providing any radical transformation of assessment
13 practices." It is evident, in the studies reviewed above as well as in other publications in
14 international academic journals, that various terms are used to refer to computer-based
15 language testing, including: computer-adaptive, computer-aided, computer-assisted,
16 computer-enhanced, computer-mediated, computer-supported, and technology-enhanced.
17 All these terms imply that computers play a peripheral role, as a supporter, enhancer, or
18 mediator of communication and the demonstration of language abilities. For innovations in
19 computer-based language testing to really occur, it is imperative that researchers, assessors,
20 and educators consider computers not only as a delivery platform but also as an integrated
21 part of the language construct to be assessed. Following this notion, hereafter we
22 recommend using the term computer-integrated^{viii} instead of computer-based. However,
23 there is a sensitive balance to strike. As Milanovic (2013, p. 32) put it, "we must try to take
24 advantage of the benefits technology has to offer without the technology tail wagging the
25 learning and assessment dog." Or, in Douglas' (2000, p. 275) words, "language
26 testing...driven by technology, rather than technology being employed in the services of
27 language testing, is likely to lead us down a road best not traveled." Douglas (2013, p. 6)
28 further reminded us that "the use of technology for its own sake can lead to the trivializing
29 of language test tasks by limiting what we can include in our tests to those things that can be
30 delivered easily by computers or the Internet or that can be scored easily by machines."

31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
The reconceptualization of computer-integrated language testing will facilitate and promote
innovations in test design, especially in task formats and assessment criteria. As a result,
there needs to be a shift of research focus. There should be fewer studies on the
comparability between computer-integrated and paper-based tests or on the adverse or
beneficial impacts of computer literacy on test performance, as in the vast majority of the
previous studies in China. Instead, more studies need to focus on the comparability between
different computer technologies or platforms. If computer literacy is accepted as an
essential part of the language construct to be assessed, computer-integrated and paper-
based tests may no longer share the same level of comparability, in theory and by design, as
current computer-based and paper-based tests do. Instead, future comparability studies
may focus on comparability between different computer-delivery platforms, posing research
questions such as, How does language performance differ according to the use of certain
technologies? What has been considered as critical features of test delivery environments
and test takers' computer literacy in the past or at present may soon become outdated or
irrelevant. For example, the screen resolution of desktops as researched in Bridgeman,
Lennon and Jackenthal (2003) a decade ago is hardly a controversial issue now.

1
2
3 Computer literacy is an evolving concept, which must necessarily be a long-lasting concern
4 of test providers, given the rapid development of different computer technologies and
5 platforms (desktops, laptops, tablets, virtual worlds, etc.). For example, studies in the 1990s
6 (Powers & Potenza, 1996) and 2010s (Ling, 2016; Ling & Bridgeman, 2013) on the
7 comparability between writing on desktop and laptop computers produced conflicting
8 findings. In Powers and Potenza (1996), participants were in favour of desktop computers,
9 and essays written on desktop computers achieved higher scores than those written on
10 laptop computers. However, in Ling and Bridgeman (2013, p.118), “essays produced using a
11 laptop were comparable to those produced using desktop computers on essays’ scores,
12 lengths, and writing speed.” Ling (2016) reported that taking a test on an iPad was similar to
13 taking a test on a desktop computer for experienced users of these two types of devices.
14 Future computer technologies may become more interactive and intuitive to use.
15 Multimodal delivery of language test content (visual, audio, animations, virtual reality, etc.)
16 via computer technology, which was not possible in paper-based tests, may better represent
17 the progressively evolving construct of language use and hence improve task authenticity. In
18 this respect, it will be fruitful to promote efforts to research innovative multimodal and
19 interactive tasks, the inter-operationability of such tasks in different platforms, and their
20 differential impacts on performance of test takers of different characteristics and for
21 different assessment purposes.
22
23
24
25
26

27 It may, however, take a long time and considerable effort to reconceptualize the construct
28 of computer-integrated language testing and to shift the focus of research from
29 comparability to issues in multimodality and inter-operationability. At least three other
30 areas require immediate action and can produce more fruitful research evidence to broaden
31 the scope, depth, and quality of research on the current practices of computer-integrated
32 English language testing in China.
33
34

35 Firstly, given the stakes and the impacts that the IB-CET has on teaching, learning and
36 Chinese society in general, more high quality research studies on the IB-CET, whether
37 independent from or commissioned by the CET, are urgently needed. The IB-CET is the
38 largest computer-based high-stakes English test in China, however, there are only a small
39 number of research publications on the IB-CET in Chinese academic journals or elsewhere;
40 these studies tended to be small-scale and conducted prior to 2011. There does not seem to
41 be any research publication on the automated evaluation system used for marking the IB-
42 CET speaking task performance. This lack may be because the IB-CET is a live, consequential
43 test and therefore for security reasons data about the test or its tasks are not released
44 beyond the National CET Committee. Nevertheless, there should be more publications
45 evaluating the IB-CET, such as Jin and Yan’s (this issue), if the Committee can release data for
46 research purposes.
47
48
49
50

51 Secondly, studies on the comparability of cognitive processes among examinees taking
52 computer-based and paper-based tests would be a welcome addition to current knowledge
53 about the effects of the two delivery modes. Almost all the Chinese empirical studies on the
54 comparability between computer-based and paper-based tests, or the impacts of computer
55 literacy on test performance, have relied primarily on test results as research data, except
56 for Jin and Yan’s (this issue) investigations of test-taking processes. More research is needed
57
58
59
60

1
2
3 to investigate the comparability between the two delivery modes in terms of test-taking
4 cognitive processes at an individual as well as at group-level (Yu, 2010). Taking computer
5 literacy as an example, group-level analyses might have demonstrated that students' test
6 scores were not affected by their computer literacy, however, at individual level, computer
7 literacy might well affect certain test taker's cognitive process (Yu, 2010). Computer
8 technology provides ample opportunities and data to do this kind of research. Test takers'
9 response time, keyboarding speed, confidence level, and test taking efforts (Setzer, Wise,
10 van den Heuvel, & Ling, 2013), just to name a few sources of data, can be readily recorded.
11 Streamlining computer-based language tests with eye-tracking devices provides further
12 opportunities to record students' eye movements as an indicator of their attentional and
13 test-taking processes (see Yu, He & Isaacs, in press). Analyses of test takers' cognitive
14 processes can help not only to understand the validity of the tasks but also to deter and
15 detect cheating or task-irrelevant behaviours during a test. Preventing cheating and
16 enhancing test security have been one of the motives for creating the current IB-CET (see Jin
17 & Yan, this issue).

18
19
20
21
22 Thirdly, more independent, transparent, and comparative research on the quality of
23 automated evaluation engines is needed to assure test takers that they are assessed fairly.
24 All the Chinese empirical studies to date have reported how well their automated evaluation
25 systems predicted human scores; however, as Carr (2014) rightly pointed out, research on
26 automated evaluation systems "has been conducted by the companies developing the
27 systems, and...there is a marked lack of independent research comparing different systems
28 head to head." This limitation also applies to the existing Chinese studies. Independent
29 research is needed to advance technological breakthroughs as well as transparency. In
30 addition to more rigorous and independent research, it is equally important to expand the
31 focus of research on automated scoring. There are a number of high priority topics that have
32 hardly been explored, such as how test takers interact with tasks that use automated scoring,
33 what test-taking processes and strategies appear, how score users (e.g., university admission
34 tutors and language support staff) interpret and use the test scores assigned by automated
35 scoring engines, and the impact of the use of automated scoring on language teaching,
36 learning, and test preparation. Future automated evaluation engines should build not only
37 on what computers can do but also on what the construct of computer-integrated
38 communication or language performance should be.
39
40
41
42
43
44
45

46 Conclusion

47
48
49 This article provides a snapshot of the current research and practice on computer-integrated
50 English language testing in China as conducted by Chinese scholars and published in major
51 Chinese academic journals. The quality of these publications is not at the same level as
52 appears in Jin and Yan (this issue) and He and Min (this issue); nevertheless, the themes and
53 focuses of these studies help to identify three key areas of current endeavours in
54 researching and using computer-based English language testing in China: computer-adaptive
55 testing, the national IB-CET and certain institutional achievement tests at universities, and
56 automated evaluation systems for speaking, writing, and translation assessment.
57
58
59
60

1
2
3 Considering these research focuses and findings, two areas should shape future agendas for
4 research. Firstly, the construct of language use in computer-based tests needs
5 reconceptualizing by integrating computer technology not only as a delivery platform but
6 also as an integral component of the language construct to be assessed. This innovation
7 should facilitate and promote innovations in test design, especially in task formats and
8 assessment criteria. Secondly, as a consequence, comparability studies should shift their
9 research focuses from studying the comparability between computer-based and paper-
10 based tests to studying the comparability between different computer technologies and
11 platforms, the inter-operationability of innovative multimodal and interactive tasks in
12 different delivery platforms, and their potential impact on the performance of test takers of
13 different characteristics for different assessment purposes. However, given the current
14 status of research on computer-based English language testing in China three more pressing
15 issues require immediate action: (a) more high quality research on the validity of the IB-CET
16 and its automated evaluation of speaking task performance, (b) more high quality research
17 on students' test taking cognitive processes, and (c) more independent, transparent, and
18 comparative research on the quality of automated evaluation engines, which should be
19 based on the construct of computer-integrated testing, rather than on the construct of
20 traditional paper-based testing or communication. The findings from these studies will have
21 implications beyond Chinese contexts in promoting better use of computer technologies in
22 and for language assessment.
23
24
25
26
27
28
29

30 Acknowledgements

31 The Editors of this special issue—Alister Cumming and David Qian—and Professor Lianzhen
32 He of Zhejiang University during her visit as Benjamin Meaker Visiting Professor at the
33 University of Bristol provided insightful comments and feedback on an earlier draft of this
34 article. This article was first presented as a keynote speech at the inaugural conference of
35 the Asian Association of Language Assessment in October 2014 in Hangzhou, China.
36
37
38
39

40 References

- 41
42
43 Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen
44 resolution, and display rate on computer-based test performance. *Applied*
45 *Measurement in Education*, 16(3), 191-205.
- 46 Cai, J. (2005). Validity, reliability and practicality of computer-based oral proficiency test.
47 *Foreign Language World*, 26(4), 66-75. [蔡基刚. (2005). 大学英语四、六级计算机口
48 语测试效度, 信度和可操作性研究. *外语界*, 26(4), 66-75.]
- 49 Cai, J. & Wang, Z. (2009). A study of validity and reliability of Internet-based English writing
50 testing *Foreign Language World*, 30(3), 52-58. [蔡基刚, & 汪中平. (2009). 英语网考
51 的写作效度和信度研究. *外语界*, 30(3), 52-58.]
- 52 Carr, N. T. (2014). Computer-automated scoring of written responses. In A. Kunnan (Ed.), *The*
53 *Companion to language assessment*. Malden, MA: John Wiley & Sons, Inc.
54 DOI: 10.1002/9781118411360.wbcla124
55
56
57
58
59
60

- 1
2
3 Chalhoub-Deville, M. B. (2010). Technology in standardized language assessments. In R.
4 Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed.) (pp. 511-526).
5 Oxford: Oxford University Press.
- 6 Chapelle, C. A. (2010). Technology in language testing. In Fulcher, G. & Trasher, R. *Language*
7 *testing videos*. In association with ILTA. Available: <http://languagetesting.info>.
- 8 Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*.
9 Cambridge, UK: Cambridge University Press.
- 10 Chen, H. (2009). A proposal on the verification model of the validity equivalence between
11 PBLT and CBLT. *Foreign Language World*, 30(3), 73-80 [陈慧麟. (2009). 基于纸笔的
12 语言测试和基于计算机的语言测试之间效度对等性验证模式初探. *外语界*, 30(3),
13 73-80.]
- 14 Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*,
15 25(1), 15-37.
- 16 Dai, Z. (2011). A study of the reliability of computerized oral proficiency interview.
17 *Computer-Assisted Foreign Language Education*, 33(2), 45-50. [戴朝晖. (2011). 计算
18 机口语考试信度研究. *外语电化教学*, 33(2), 45-50.]
- 19 Dai, Z. & You, Q. (2010). Multi-facets Rasch Model analysis of rater bias in Computerized EFL
20 Oral Proficiency Interview. *Foreign Language World*, 31(5), 87-95. [戴朝晖, & 尤其
21 达. (2010). 大学英语计算机口语考试评分者偏差分析. *外语界*, 31(5), 87-95.]
- 22 Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, UK: Cambridge
23 University Press.
- 24 Douglas, D. (2013). Technology and language testing. In C. A. Chapelle (Ed.), *The*
25 *Encyclopedia of Applied Linguistics*. Malden, MA: Wiley-Blackwell.
26 DOI: 10.1002/9781405198431.wbeal1182
- 27 Du, J. & Gui, S. (2000). An experimental study of computerized diagnostic testing of reading.
28 *Foreign Language Teaching and Research*, 32(5), 345-351. [杜金榜, & 桂诗春.
29 (2000). 电脑化阅读诊断测试的实验研究. *外语教学与研究*, 32(5), 345-351.]
- 30 Gao, B. (2007). A comparative study of COPT and DOPT. *Computer-Assisted Foreign*
31 *Language Education*, 29(2), 73-76. [高丙梁. (2007). 计算机口试与面试的比较研究.
32 *外语电化教学*, 29(2), 73-76.]
- 33 Ge, S. (2010). A comparative study of automated essay scoring techniques for college
34 students' English writing. *Journal of Guangdong University of Foreign Studies*, 21(3),
35 87-90. [葛诗利. (2010). 大学英语作文自动评分方法比较研究. *广东外语外贸大学*
36 *学报*, 21(3), 87-90.]
- 37 Ge, S. & Chen, X. (2009). Cluster analysis of college English writing in automated essay
38 scoring. *Computer Engineering and Applications*, 45(6), 145-148. [葛诗利, & 陈潇潇.
39 (2009). 文本聚类在大学英语作文自动评分中应用. *计算机工程与应用*, 45(6),
40 145-148.]
- 41 He, L. (1999). Designing cognitive computer-adaptive tests. *Modern Foreign Languages*,
42 22(2), 169-183. [何莲珍. (1999). 认知计算机适应性考试模型的设计. *现代外语*,
43 22(2), 169-183.]
- 44 Huang, M. & Qin, C. (2009). An investigation into test takers' adaptability to the Internet-
45 based CET. *Foreign Language World*, 30(5), 90-96. [黄敏, & 覃朝宪. (2009). 全国大
46 学英语网络考试考生情况调查——以适应性为研究视角. *外语界*, 30(5), 90-96.]
- 47 Huang, Y. & He, L. (2013). Approach to fitting testlet for computerized adaptive language
48 testing. *Computer-Assisted Foreign Language Education*, 35(2), 29-34. [黄妍, & 何莲
49 珍. (2013). 计算机自适应语言测试的题组拟合方法. *外语电化教学*, 35(2), 29-34.]
- 50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Jiang, J. (2013). An automatic approach to evaluating the linguistic quality of English-Chinese
4 translations. *Modern Foreign Languages*, 36(1), 85-91. [江进林. (2013). 英译汉语言
5 质量自动量化研究. *现代外语*, 36(1), 85-91.]
- 6
7 Jiang, J. & Wen, Q. (2010). A comparative study of N-gram and translation unit alignment in
8 automated scoring of students' English-Chinese translation. *Modern Foreign*
9 *Languages*, 33(2), 177-184. [江进林, & 文秋芳. (2010). N 元组和翻译单位对英译汉
10 自动评分作用的比较研究. *现代外语*, 33(2), 177-184.]
- 11
12 Jiang, J. & Wen, Q. (2012). Computer scoring models for EFL learners' English-Chinese
13 translation in large-scale tests. *Computer-Assisted Foreign Language Education*,
14 34(2), 3-8. [江进林, & 文秋芳. (2012). 大规模测试中学生英译汉机器评分模型的
15 构建. *外语电化教学*, 34(2), 3-8.]
- 16
17 Jin, L. (2011). A brief study of computer-aided test of oral English. *Foreign Language and*
18 *Literature*, 27(4), 126-130. [金力. (2011). 计算机辅助大学英语口语测试研究. *外国*
19 *语文*, 27(4), 126-130.]
- 20
21 Jin, Y., & Wu, J. (2010). A preliminary study of the validity of the Internet-based CET-4—
22 Factors affecting test-takers' perception of and performance on the test. *Computer-*
23 *Assisted Foreign Language Education*, 32(2), 3-10. [金艳, & 吴江. (2010). 大学英语
24 四级网考效度初探——影响考生评价和考试成绩的因素分析. *外语电化教学*,
25 32(2), 3-10.]
- 26
27 Li, M., Yang, X., Feng, G., Wu, M., Chen, J., & Hu, G. (2008). Machine scoring of reading aloud
28 item of large-scale college English oral tests. *Foreign Language World*, 29(4), 88-95.
29 [李萌涛, 杨晓果, 冯国栋, 吴敏, 陈纪梁, & 胡国平. (2008). 大规模大学英语口语测
30 试朗读题型机器阅卷可行性研究与实践. *外语界*, 29(4), 88-95.]
- 31
32 Li, X. & Liu, J. (2013). Ensemble learning based essay automated scoring algorithm for
33 Chinese English learners. *Journal of Chinese Information Processing*, 27(5), 100-106.
34 [李霞, & 刘建达. (2013). 适用于中国外语学习者的英文作文全自动集成评分算法.
35 *中文信息学报*, 27(5), 100-106.]
- 36
37 Li, Y. (2009). An empirical study of the effect of the large-scale computer-assisted Spoken
38 English Test. *Foreign Language World*, 30(4), 69-76. [李玉平. (2009). 大规模计算机
39 辅助英语口语测试效果实证研究. *外语界*, 30(4), 69-76.]
- 40
41 Li, Y. & Ge, S. (2008). The validity of word list in automated essay scoring for college students.
42 *Foreign Languages and Their Teaching*, 24(10), 48-52. [李艳, & 葛诗利. (2008). 大学
43 英语作文自动评分中分级词表的效度研究. *外语与外语教学*, 24(10), 48-52.]
- 44
45 Ling, G. (2016). Does it matter whether one takes a test on an ipad or a desktop computer?
46 *International Journal of Testing*, 16(4), 352-377. doi:
47 10.1080/15305058.2016.1160097
- 48
49 Ling, G., & Bridgeman, B. (2013). Writing essays on a laptop or a desktop computer: Does it
50 matter? *International Journal of Testing*, 13(2), 105-122.
51 DOI: 10.1080/15305058.2012.690012
- 52
53 Liu, P. (2011). Investigation on the problems of Internet-based CET 4 & 6 and solutions from
54 test-takers' perspective. *Modern Educational Technology*, 21(12), 77-81. [刘萍.
55 (2011). 考生视角下大学英语四、六级网考的问题与对策. *现代教育技术*, 21(12),
56 77-81.]
- 57
58 Liu, Z. & Liu, D. (2015). The design and implementation of a system for the automatic
59 assessment of learners' translations. *Journal of PLA University of Foreign Languages*,
60 38(2), 109-115. [刘泽权, & 刘鼎甲. (2015). 学习者英译文自动评估系统的设计与实
实现. *解放军外国语学院学报*, 38(2), 109-115.]
- Milanovic, M. (2013). A look into the future. *Research Notes* (51), 31-33.
- Powers, D. E., & Potenza, M. (1996). *Comparability of testing using laptop and desktop computers* (ETS RR-96-15). Princeton, NJ: Educational Testing Service.

- 1
2
3 Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment:
4 Affective effects on test takers. *Language Assessment Quarterly*, 6(2), 1-13.
- 5 Qiao, H., Dong, B., & Liu, C. (2012). A study on automated scoring of PETS computer-based
6 speaking test. *Foreign Language Testing and Teaching*, 2(3), 47-52. [乔辉, 董滨, & 刘
7 常亮. (2012). PETS 计算机辅助口试自动评分技术研究. *外语测试与教学*, 2(3), 47-
8 52.]
- 9 Qiu, D., Ji, P., Wan, J. & Cheng, Y. (2005). A study on computer-based listening and speaking
10 tests for college students. *Foreign Language World*, 26(4), 76-79 . [邱东林, 季佩英,
11 万江波, & 程寅. (2005). 大学英语听说机考尝试. *外语界*, 26(4), 76-79.]
- 12 Sawaki, Y. (2012). Technology in language testing. In G. Fulcher & F. Davidson (Eds.), *The*
13 *Routledge handbook of language testing* (pp. 426-437). Abingdon, UK: Routledge.
- 14 Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An Investigation of examinee
15 test-taking effort on a large-scale assessment. *Applied Measurement in Education*,
16 26(1), 34-49. doi: 10.1080/08957347.2013.739453
- 17 Si, Y. (2008). An empirical study of large-scaled computer-assisted diagnostic testing of
18 spoken business English. *Computer-Assisted Foreign Language Education*, 30(1), 67-
19 71 . [司耀龙. (2008). 基于计算机的大规模商务英语口语诊断测试实践研究. *外语*
20 *电化教学*, 30(1), 67-71.]
- 21 Tang, J. & Liu, X. (2009). Effects of delivery mode on test performance – A comparative study
22 on computer-based and paper-based tests. *Distance Education in China*. 唐锦兰, &
23 刘晓悦. (2009). 考试媒介对于考生成绩的影响研究——一项英语机考与纸笔考
24 试成绩对比分析. *中国远程教育*(5), 57-61.
- 25 Wang, L. & Chang, B. (2009). Research on the human-aided auto-assessment for translation
26 tests in College English. *Computer-Assisted Foreign Language Education*, 31(4), 17-
27 21. [王雷, & 常宝宝. (2009). 大学英语翻译考试人工辅助计算机评分初探. *外语*
28 *电化教学*, 31(4), 17-21.]
- 29 Wang, Y. (2004). A study of online marking of CET compositions. *Foreign Language World*,
30 25(5), 74-79. [王跃武. (2004). 大学英语四、六级考试作文网上阅卷实验研究. *外语*
31 *界*, 25(5), 74-79.]
- 32 Wang, Y., Zhu, Z., Yang, H. (2006). The implementation of a many-facet Rasch measurement
33 to the reliability estimates of online marking *Foreign Language World*, 27(1), 69-76 .
34 [王跃武, 朱正才, & 杨惠中. (2006). 作文网上评分信度的多面 Rasch 测量分析. *外*
35 *语界*, 27(1), 69-76.]
- 36 Wang, Y., Zhu, Z., Yang, H. (2006). The implementation of a many-facet Rasch measurement
37 to the reliability estimates of online marking *Foreign Language World*, 27(1), 69-76 .
38 [王跃武, 朱正才, & 杨惠中. (2006). 作文网上评分信度的多面 Rasch 测量分析. *外*
39 *语界*, 27(1), 69-76.]
- 40 Wen, Q. , Qin, Y., & Jiang, J. (2009). Application of bilingual alignment technology to
41 automatic translation scoring of English test. *Computer-Assisted Foreign Language*
42 *Education*, 31(1), 3-8 . [文秋芳, 秦颖, & 江进林. (2009). 英语考试翻译自动评分中
43 双语对齐技术的应用. *外语电化教学*, 31(1), 3-8.]
- 44 Xu, Z. , Xie, X., Liu, C. , Chen, X., Liu, F., & Gu, J. (2013). An empirical study on large-scale
45 computer-assisted college oral English test. *Modern Educational Technology*, 23(8),
46 76-80. [徐智鑫, 谢小苑, 刘长江, 陈向俊, 刘芳, & 谷健飞. (2013). 高校大规模计算
47 机辅助英语口语测试实证研究. *现代教育技术*, 23(8), 76-80.]
- 48 Yan, K., Hu, G., Wei, S., Dai, L. , Li, M., Yang, X., & Feng, G. (2009). Automatic evaluation of
49 English retelling proficiency in large machine based oral English tests. *Journal of*
50 *Tsinghua University (Science and Technology)*, 49(S1), 1356-1362. [严可, 胡国平, 魏
51 思, 戴礼荣, 李萌涛, 杨晓果, & 冯国栋 (2009). 面向大规模英语口语机考的复述题
52 自动评分技术研究. *清华大学学报: 自然科学版*, 49(S1), 1356-1362.]
- 53 Yan, K., Hu, G., Wei, S., Li, M., Yang, X. & Feng, G. (2010). A primary study on computerised
54 automatic marking of English recitation proficiency. *Computer Applications and*
55 *Software*, 27 (7), 164-168. [严可, 胡国平, 魏思, 李萌涛, 杨晓果, & 冯国栋 (2010).
56 计算机用于英语背诵题的自动评分技术初探. *计算机应用与软件*. 27 (7), 164-168.]
- 57
58
59
60

- 1
2
3 Yang, Y. & Li, M. (2010). A study on attitudes of college students in the computer-based oral
4 English test environment. *Foreign Language World*, 31(6), 78-84. [杨艳霞, & 李萌涛.
5 (2010). 大学英语口语机考环境下大学生计算机使用态度研究. *外语界*, 31(6), 78-
6 84.]
- 7 Yin, N., Zheng, Y., Wang, L., & Xin, D. (2010). A Comparative study on the effects of COPT
8 and OPI on oral fluency. *Computer-Assisted Foreign Language Education*, 26(3), 25-
9 29. [尹楠, 郑玉荣, 王丽丽, & 辛丹. (2010). 机辅与面试对口语流利性影响的对比
10 研究. *外语与外语教学*, 26(3), 25-29.]
- 11 Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of
12 extended texts. *Language Assessment Quarterly*, 7(2), 119-136.
- 13 Yu, G., He, L., & Isaacs, T. (in press). *The cognitive processes of taking IELTS Academic writing*
14 *task one: An eye-tracking study*. Report to British Council/Cambridge Assessment.
- 15 Zeng, Y. (2002). A preliminary study on individualized self-adaptive testing. *Foreign*
16 *Language Teaching and Research*, 34(4), 278-282. [曾用强. (2002). 个性化自适应性
17 测试探索. *外语教学与研究*, 34(4), 278-282.]
- 18 Zeng, Y. (2010). The Computerized Oral English Test of the National Matriculation English
19 Test. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese*
20 *learner* (pp. 234-247). Abingdon, UK: Routledge.
- 21 Zhang, S., & Yu, P. (2010). Online rater training for CET4 writing assessment. *Foreign*
22 *Language World*, 31(5), 79-86. [张森, & 于朋. (2010). 大学英语四级考试作文网上
23 评阅信度保障研究. *外语界*, 31(5), 79-86.]
- 24 Zhang, W. (1999). An experiment with self-adaptive testing. *Journal of PLA University of*
25 *Foreign Languages*, 22(3), 53-55. [张武保. (1999). 自适应性测试的实验研究. *解放*
26 *军外国语学院学报*, 22(3), 53-55.]
- 27 Zhou, Y. & Zeng, Y. (2016). Many-facet Rasch model analysis of computer automatic scoring
28 in a computer-based English listening-speaking test. *Foreign Language Testing and*
29 *Teaching*, 6(1), 22-31. [周燕, & 曾用强. (2016). 机助英语听说考试计算机自动评分
30 的多层面 Rasch 模型分析. *外语测试与教学*. 6(1), 22-31.]
- 31 Zhu, Y. & Zhang, X. (2009). An exploration of practice on the computer-based testing in
32 College English. *Computer-Assisted Foreign Language Education*, 31(2), 63-67. [朱音
33 尔, & 张肖莹. (2009). 基于网络的大学英语机考探索与实践. *外语电化教学*, 31(2),
34 63-67.]

Notes

ⁱ www.cnki.net; the database houses full-texts of all current Chinese journals (from 1915 onwards).

ⁱⁱ Before the CET was computerized, efforts to use computer technology in CET mainly focused on “online” marking of writing for the paper-based CET in order to improve and monitor marking reliability and efficiency. The writings were scanned to be marked. A few CET-sponsored studies reported the benefits of using online marking over “conference marking” thanks to the real-time monitoring function of the online marking system (Wang, 2004; Wang, Zhu & Yang, 2006; Zhang & Yu, 2010). Although strictly speaking, these studies were not about computer-based tests, the current practice of online marking of writings produced in computer-based CET has been influenced by the findings of these studies, and therefore, we think they are worth mentioning here.

ⁱⁱⁱ ISBN of the product: 978-7-900717-85-6/H-53; Listed Price: 200k Chinese Yuan;
http://www.ssit.cc/product_in.aspx?PID=24&CategoryName=%E5%A4%A7%E5%AD%A6%E8%8B%B1%E8%AF%AD&CID=1;
The Test System was developed in collaboration with the University of Science and Technology of China.

^{iv} <http://www.iflytek.com/en/index.html>. The system has an automated evaluation component.

^v <http://www.gzlang.com/paperless-examination.aspx>

^{vi} <http://www.wingsoft.com.cn/product3.jsp>

^{vii} This university developed iflytek which contains automated evaluation of speaking performances, so it is possible that it was these researchers who developed the automated evaluation system in iflytek.

^{viii} However, this proposal is not to suggest that all of these terms should be replaced by just one term, i.e., computer-integrated assessment, as each term tends to have specific meanings and operate in different assessment contexts. In cases where the use of computer technology is deemed to be part of the construct of the language task/test, it is more appropriate to use the term “computer-integrated” assessment.

For Peer Review Only