On the validity and reliability of a computer-assisted English speaking test

Zhihong Lu[1], Zhenxiao Li[2], Leijuan Hou[3]
Beijing University of Posts and Telecommunications
Beijing, China
luzhihong@bupt.edu.cn[1], briannalee@163.com[2], houleijuan2008@sina.com[3]

***Abstract*—Ever since the emergence of computer in 1950s and its application in many practical areas of science and educational disciplines, Computer-assisted Language Learning (CALL) has been gradually implemented to the field of English as a Foreign Language (EFL) teaching, and a computer-based speaking test has become an indispensable component of an EFL integrated language skill test. Experts and scholars from both computer application technology and EFL instructors have conducted a certain number of studies concerning computer-assisted English speaking test. Since validity and reliability are the most essential elements of a high quality language test, based on the communicative language teaching approach, an experiment at the first author's English Audio-video Speaking Course (EAVSC) in a digital language lab was carried out to investigate the usefulness of a computer-assisted English speaking test within the framework of Communicative Language Testing (CLT). Through a series of statistical analyses with SPSS software, it can be concluded that the validity and reliability of the computer-assisted English speaking test can be achieved within the framework of CLT. It also analyzes the backwash of the test format as well as its pedagogical implications in further related studies.**

***Keywords*—CALL; validity; reliability; English speaking test; EFL teaching; CLA; CLT**

I. Introduction

It is widely acknowledged that the major problem existing in Chinese EFL learners lies in their listening and speaking competence, especially their speaking ability. It has been highly emphasized from *College English Curriculum Requirements* that college English teaching should adopt "the computer- and the classroom-based multimedia teaching model" (Department of Higher Education, 2007). Teaching objectives should shift from English reading skills to listening and speaking abilities, with much emphasis on Communicative Language Ability (CLA) and autonomous learning ability. To achieve the goal, various measures have been taken in the field and EFL instructors and language test designers shoulder the responsibility for teaching and testing if EFL learners' listening and speaking abilities have been improved under such newly adopted teaching model.

Integrating modern information and communication technologies (ICT) with EFL teaching and learning, a computer-assisted English speaking test can provide a convenient, efficient, and reliable approach to assess the individual learner's spoken skill. Whatever form an English speaking test may take, the most essential element to be considered should be its validity and reliability, which serve as the core of an effective test.

II. Literature Review

A. Communicative Language Testing Model

Lyle F. Bachman's CLA model has been regarded as the most influential framework of its kind and laid a solid foundation for the birth of communicative language testing model [1]. It consists of three parts: language competence, strategic competence and psycho-physiological mechanisms.

Language competence comprises essentially a set of specific knowledge components that are utilized in communication via language [2]. It includes organizational competence, which consists of grammatical and textual competence, and pragmatic competence, which consists of illocutionary and sociolinguistic competence. Strategic competence is seen as the capacity that relates language competence, or knowledge of language, to the language user's knowledge structures and the features of the context in which communication takes place. Psycho-physiological mechanisms involved in language use characterize the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented [2].

B. English Speaking Test and CALL

As an indispensable part of communicative language testing module, English speaking tests can be categorized into three forms: the direct speaking test, the half-direct speaking test and the indirect speaking test [3]. Each has its advantages and disadvantages and plays a positive role in certain historical periods or in specific teaching environments. The Business English Certificate (BEC) spoken test, the interview session of the IELTS and the Test of Spoken English (TSE) respectively are three international representatives of the above test types [3]. Various changes have taken place in speaking test content in accordance with the shift of traditional testing module to communicative testing. Tests are designed with a focus on assessing EFL learners' ability to use the target language dynamically in real-life situations rather than artificial simulations.

Computer-assisted Language Learning (CALL) is, as Ken Beatty put it, any process in which a learner uses a computer and, as a result, improves his or her knowledge.

This study is part of the humanities and social sciences project "Research on Multidimensional Assessment for a Web-based English Audio-video Speaking Course" (12YJA740052), supported by the Ministry of Education in China.

CALL has become increasingly integrated into research and practice in the general skills of reading, writing, speaking and listening [4].

Researchers both home and abroad have done studies in respect of CALL and speaking tests. Warschauer and Healey summarized the development of CALL into three phrases: Behaviouristic, Communicative and Integrative CALL [5]. Different from Warschauer and Healey's terms which imply a historical division, Bax's terminologies of "Restricted CALL" and "Open CALL" were put forward from the perspective of 'approaches' [6]. In China, related research has also been carried out. Han gave a full view of Bachman's CLA in a series of articles [7], while Zou explored the authenticity of different types of oral tests [8]. In the empirical area, Cai analyzed the validity, reliability and practicality of computer-based oral proficiency tests, taking the national College English Test-Spoken English Test (CET-SET) for an example [9]. Li analyzed the large-scale computer-assisted SET from the perspective of humanistic concepts and provided some encouraging insights accordingly [10].

Although substantial contributions have been made by previous scholars, the number of empirical studies regarding computer-assisted SET is still relatively small. In particular, relative research on learners' CLA is insufficient.

C. *Validity and Reliability*

According to Cohen, validity refers to whether the assessment instrument actually measures what it purports to measure; the reliability of a test asks whether an assessment instrument administered to the same respondents a second time would yield the same results [11]. Hughes considered that the validity of a test includes content validity, criterion-related validity, construct validity and face validity; while the reliability of a test is usually approached from two angles: the reliability of the test paper and scorer reliability [12].

Integrating computer-assisted SET with Bachman's CLT model, this paper aims to explore the validity and reliability of English speaking tests at the first author's English Audio-video Speaking Course (EAVSC).

III. RESEARCH DESIGN

A. *Research Objectives*

Through both qualitative and quantitative analyses of correlated data from the EAVSC, the research was designed to address the following questions:

- To what extent is a computer-assisted SET valid and reliable within the framework of the CLT model?
- To what extent can such a SET format reflect EFL learners' real speaking ability?

B. *Subjects*

The research was based on the first author's EAVSC and the teaching experiment lasted for one semester from September 2013 to January 2014. The participants were 34 juniors of non-English major students. Most of them have studied English for approximately eight years. All of them have passed CET-4 with a score above 545 (The total score is 710.).

C. *Teaching Context*

The two classes were conducted in a digital language lab, equipped with a digital learning system, WE-LL6000, which makes it possible for synchronous computer-mediated oral communication, random grouping, speech recording, and the Internet connections. When assigning tasks, the whole class can be divided into several groups or pairs at random or by set rules. Students can talk to their desk partners or group members without face-to-face contact, eliminating the possible shyness or embarrassment experienced in real life situations.

The testing environment in which students took final exams was the same as the teaching environment. In this sense, the nature of computer-assisted SET format in this study can be regarded as a semi-direct speaking test. The format along with its content could make it possible for computer-mediated interpersonal interactions. The guiding principle throughout the entire teaching procedure is to bring students into active and competent participation in various speaking activities.

D. *Experiment Instrument*

1) *Test specifications*

To assess if a test has content validity, the authors used a type of test specification to compare the subtests' relevance to the test objectives. Adopting Bachman's CLA model as a general framework, the authors made some adaptations when devising the test specifications according to the specific teaching context in this study. It is made up of three parts: the organizational competence, the pragmatic competence and the strategic competence.

2) *Pre- and post-tests*

The pre- and post-tests, which included three parts, were designed in the same format but with different topics and under the same condition. The first part was a six-minute three-person group discussion, with the score value of 30%. During the group work, each student took turns to chair one topic discussion to make it possible for each student's engagement in discussions by applying some concerning communicative strategies introduced by the instructor. The second part was a five-minute pair-work conversation between desk partners on either of the two given topics (vacation plan or college life), with the score value of 50%. In this part, students were encouraged to use both verbal and non-verbal communication strategies. The third part of the test was a one-minute personal statement task on the topic previously talked about, with the score value of 20%. The aim of the tests was to evaluate students' progress in their oral activities, therefore, the topics were suggested to be kept relevant.

Students' final SET scores were used to investigate both the test's internal and external reliability and construct validity. The pre-test SET scores were established as a criterion to measure the test's construct validity. All of the students' speaking activities were recorded for our research and the analysis through the WE-LL6000 system equipped in the lab.

E. Data Collection

The data of the pre- and post-tests were obtained by the average score of two teachers' ratings on students' recordings. The assessment criteria were constructed on the basis of two influential existing SET criteria: IELTS Speaking Band Descriptions and CET-SET. To access the face validity of the test, a certain number of sample recordings were randomly selected and transcribed into text forms. All the data were processed through SPSS 19.0.

- The discourse analysis research method was adopted to check the face validity of the tests.
- To obtain content validity, the sub-tests and the test specifications were compared to show the degree of consistency.
- To judge the construct validity, a series of Pearson Correlation Coefficient tests were employed to illustrate the correlations between sub-tests, and between sub-tests and the whole test.
- The criterion-related validity was examined through a Pearson correlation coefficient test between the scores both at the pre-SET and the post-SET in the experimental class
- A Cronbach's alpha reliability coefficient test was used to examine the internal reliability of the test.
- A paired sample *t*-test of the two teachers' ratings was employed to testify the test's external reliability.

IV. Discussion and Findings

A. Face Validity

A test is said to have face validity if it looks as if it measures what it is supposed to measure [12]. Anderson stated that instruments such as verbal report and questionnaires provide data on face validity to know students' attitudes, feedback and feelings toward SET [13]. In this study, however, in view of unavoidably involved subjective elements in the above two methods, the authors adopted a qualitative research method—discourse analysis to check if the students have fulfilled their tasks effectively.

From both the recordings and the randomly selected samples of transcript texts, it is clear that, by and large, average students performed well during the test.

In the first part, students' strategic ability was the main concern. Discussion can be regarded as an effective format to examine learners' comprehensive abilities. By analyzing the sample transcriptions, the authors have found that more than 90% of students actively participated in the group discussion. On the one hand, most of them were able to express their opinions with a certain command of language structures; on the other hand, they could skillfully apply linguistic and communicative strategies in specific contexts. There were not many or long pauses, instead, there were some continuous heated discussions among groups.

The second part of the test involving authentic tasks was to evaluate students' pragmatic competence. Students were asked to talk about some topics closely related to their lives. Furthermore, these communicative tasks had an important function, that is, to facilitate the exchange of information among students. The students were using the target language for real-life social interactions provided through the digital language lab. From their conversations, it proved that this part preferably measured their illocutionary and sociolinguistic competence, the two components of pragmatic competence.

In the third part, guided by the teacher's instruction, students were allowed to prepare the topic by taking notes for the first fifteen seconds, and then proceeded with their statements during the next forty-five seconds. Students began with a clean opinion on the topic and then provided two or three reasons to support their statements. Following Bachman's CLA, this part was designed to test students' organizational competence, containing grammatical and textual competence. The validity of the test is evident from the statistical analysis of the randomly selected samples.

B. Content Validity

Combining the models of Bachman, as well as Canale and Swain, understanding of strategic competence creates three domains which are meant to be examined in the test: one's knowledge of the world, language use context and collaborative strategies. Communication is a dynamic process, because it requires speakers and hearers to proceed with a given topic utilizing their prior strategies to improve respective English communication skills. In the process of discussions, students were asked to take turns to chair discussions on specific topics, getting the other group members involved in comments sharing. It enables everyone in the group to obtain an equal opportunity to bring his/her communicative competence into full play.

In part two, students were asked to exchange information with the target language and thus examined their illocutionary competence. Inaccuracy or ambiguity was tolerated as long as one could successfully make himself/herself understood. Language served as a medium to maintain interpersonal relationships. Socio-cultural factors were inevitably involved since students came from different regions in China. Since part two was about real-life tasks, students could relate their topics to their life experiences. Therefore, their pragmatic competence which includes illocutionary competence and sociolinguistic competence can be measured by the extent to which they fulfilled assigned communicative tasks.

The personal statement was designed to assess students' organizational competence in terms of their choices of words, appropriate forms of words, sentence structures, pronunciation, cohesion and rhetorical organization. In addition to correct selection of words and sentence patterns, they should also make their speech coherent and well organized. Therefore, the third part of the test has fully measured students' organizational competence which includes grammatical competence and textual competence, to see if they have a good mastery of basic linguistic elements of English.

C. Construct Validity

Quantitative approaches for validation include internal correlations, factor analysis, multi-trait and multi-method analysis, comparison between students' bio-data and psychological characteristics [14]. These correlation coefficients between different components are expected to be fairly low in that they are designed to evaluate learners' different traits or skills, possibly in the order of 0.30 to 0.50; while the correlation coefficients between each sub-test and the whole test might be expected, according to the classical theory, to be higher.

The inner correlations between the sub-tests and the whole test of the two sets of scores are shown in the following two tables.

TABLE I. INNER CORRELATIONS OF THE SCORES GIVEN BY TEACHER A

		Group Discussions (GD)	**Pair-work (PW)**	**Personal Statements (PS)**	**Total**
GD	Pearson Correlation	1	0.331	0.393^{*b}	0.815**
	Sig. (2-tailed)		0.056	0.022	0.000
	N	34	34	34	34
PW	Pearson Correlation	0.331	1	0.442^{**a}	0.736**
	Sig. (2-tailed)	0.056		0.009	0.000
	N	34	34	34	34
PS	Pearson Correlation	0.393**	0.442**	1	0.743**
	Sig. (2-tailed)	0.022	0.009		0.000
	N	34	34	34	34
Total	Pearson Correlation	0.815**	0.736**	0.743**	1
	Sig. (2-tailed)	0.000	0.000	0.000	
	N	34	34	34	34

a. ***Correlation is significant at the 0.01 level (2-tailed)*

b. **Correlation is significant at the 0.05 level (2-tailed)*

TABLE II. INNER CORRELATIONS OF THE SCORE GIVEN BY TEACHER B

		Group Discussions (GD)	**Pair-work (PW)**	**Personal Statements (PS)**	**Total**
GD	Pearson Correlation	1	0.447**	0.225	0.728**
	Sig. (2-tailed)		0.008	0.201	0.000
	N	34	34	34	34
PW	Pearson Correlation	0.447**	1	0.404^{*d}	0.811**
	Sig. (2-tailed)	0.008		0.018	0.000
	N	34	34	34	34
PS	Pearson Correlation	0.225	0.404^{*d}	1	0.721**
	Sig.(2-tailed)	0.201	0.018		0.000
	N	34	34	34	34
Total	Pearson Correlation	0.728**	0.811**	0.721**	1
	Sig. (2-tailed)	0.000	0.000	0.000	
	N	34	34	34	34

c. **Correlation is significant at the 0.01 level (2-tailed)*

d. ***Correlation is significant at the 0.05 level (2-tailed)*

As shown in Table Ⅰ and Table Ⅱ, abilities that the sub-tests meant to evaluate are independent of each other and consistent to the overall abilities embodied in the whole test. A high degree of construct validity has been achieved in this test. The two tables indicate that the Pearson correlation coefficients between the sub-tests and the whole tests are high, higher than 0.70 in each set of scores. It shows that at the confidence level of 0.01 (2-tailed), these partial abilities examined in the three components pertain to the overall ability examined in the whole test to a great extent. The correlation coefficients between personal statements and pair-work are 0.442 and 0.404, respectively in the two tables, and that between personal statements and group discussions are 0.393 and 0.225 respectively; while the figures between pair-work and group discussions are 0.331 and 0.447, respectively in each table. These three groups of data, no matter whether the confidence level is 0.05 or 0.01, all indicate that the inner correlation coefficients are relatively low, around 0.30 to 0.50. This confirms that in contributing to the whole test content, the three areas focus on one specific domain and measures testees' different traits. The test has been successfully designed to avoid overlap among its three areas.

D. Criterion Validity

Apart from investigating the inner validity of the test structure, it is necessary to make a comparison between the test and other parallel independent and highly dependable assessments to show if the test is also valid externally. External validity contains concurrent validity and predictive validity. The statistic frequently used to analyze them is the correlation coefficient [14]. Predictive validity was eliminated in this study since the test was not designed to make further judgments from their performances in this case. To find a set of both independent and reliable criteria is no easy task. However, a test similar both in form and content to the target test can be accepted by most people. Therefore, a SET at the beginning of the semester is necessary since it can be served as a criterion for the post-SET. The pre-test was designed in conformity with the organization of the post-SET and administered under the same digital lab. Results of the two tests were analyzed through the mean score rated by the two teachers. The descriptive statistics and correlation coefficients between the two tests are as follows.

TABLE III. DESCRIPTIVE STATISTICS OF SCORES AT THE PRE- TEST AND FIANL- SET

	Mean	Std. Deviation	N
Final SET	83.0882	3.33365	34
Pre-test	80.5441	3.49538	34

TABLE IV. PEARSON CORRELATION COEFFICIENT BETWEEN THE PRE-TEST AND THE FINAL- SET

		Pre-test	**Final SET**
Pre-test	Pearson Correlation	1	0.870^{**e}
	Sig. (2-tailed)		0.000
	N	34	34
Final SET	Pearson Correlation	0.870^{**}	1
	Sig. (2-tailed)	0.000	
	N	34	34

e. **Correlation is significant at the 0.01 level (2-tailed)

As shown in Table Ⅲ, this test possesses external validity in the specific teaching environment. Though the means of the two sets of scores diverge slightly, their standard deviations are rather close, which implies that the discrepancy is not so significant.

As shown from Table Ⅳ, the correlation coefficient between pre-test and final SET reaches as high as 0.870. Thus, we have 99% of certainty to say that the scores between pre-test and final SET are consistent with each other. They present a similar degree of mirroring students' communicative abilities. With the same students, the same test environment and the same way of scoring as the prerequisites, students' performances in the pre-test can be well used as a criterion.

E. *Reliability*

Bachman stated that although validity is the most important quality of test scores, reliability is necessary to obtain it. Three influential models of measurement theory on reliability are: Classic true score model, Generalizability theory, and Item response theory [4]. The reliability of a test can be approached from both internal consistency and external consistency.

1) *Internal reliability*

Internal reliability refers to how consistent test takers' performances on the different parts of the tests are with each other [4]. Within the Classic true score measurement model, there are two methods to estimate internal consistency: an estimate based on correlation (Spearman-Brown split half estimate), and estimates based on item variances (the Guttman split half, the Kuder-Richardson formulae, and coefficient alpha). For a test or a questionnaire, "Cronbach's alpha is a useful coefficient for assessing internal consistency" [15]. It has served as a dependable norm in the experimental studies in social sciences. The formula is as follows:

$$\alpha = \left[\frac{k}{k-1}\right]\left[1-\frac{\sum S_i^2}{S_T^2}\right] \qquad (1)$$

Cronbach's alpha ranges from 0.00 to 0.10. The higher the coefficient is, the more reliable the test is. A perfect coefficient is 1.0. Related statistics are shown in Table Ⅴ and Table Ⅵ.

TABLE V. RELIABILITY STATISTICS—CRONBACH'S ALPHA

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
0.838	0.900	4

TABLE VI. RELIABILITY STATISTICS—ITEM-TOAL STATISTICS

	Group Discussions	Pair-work	Personal Statements	Total
Scale Mean if Item Deleted	118.703	137.574	146.444	80.544
Scale Variance if Item Deleted	32.282	34.790	36.852	12.282
Corrected Item-total Correlation	0.733	0.826	0.704	1.000
Squared Multiple Correlation	1.000	1.000	1.000	1.000
Cronbach's Alpha if Item Deleted	0.784	0.786	0.821	0.768

From Table Ⅵ it can be shown that this test has a high degree of internal reliability. According to Cohen [14], a reliability coefficient of at least 0.70 is necessary for classroom testing. In basic research, however, only when the reliability coefficient reaches 0.80 or higher can the hypothesis be accepted; as for tentative studies, the hypothesis can be accepted given that the Cronbach's alpha is higher than 0. 7. A value between 0.70 and 0.98 indicates that the test score has a high degree of reliability, while a test is said to have low reliability if the value is lower than 0.35. The Cronbach's alpha value for this study is 0.838, and if the three items are deleted, it will be 0.784, 0.786, and 0.821 respectively. Thus this test achieves internal consistency, regarding testing candidates' different aspects of language abilities via different types of sub-tests.

2) *External reliability*

Apart from investigating the internal reliability of a test, it is also essential to see if it achieves external consistency. The external reliability of a test can be examined through the level of agreement among raters. It is possible to quantify the inter-rater reliability by means of computing the Pearson correlation coefficient between the scores given by different raters. A high correlation coefficient signifies that the raters hold a similar attitude towards the candidates' performance. Another way to examine the relationship between two teachers' ratings is to compare the statistical values such as mean, standard deviation, and standard error mean. To be more exact, a paired *t*-test is used to demonstrate if there is a significant difference between different sets of results by different raters.

TABLE VII. RELIABILITY STATISTICS—PAIRED SAMPLES CORRELATIONS

	N	Correlation	Sig.
Pair 1 Teacher A Teacher B	34	0.845	0.000

As shown in the scatter gram Table Ⅶ, the Pearson correlation coefficient of the two raters is 0.845 (sig=0.000), indicating that there is strong positive correlation between the two teachers' ratings. It demonstrated that students' scores obtained are reliable and stable, which indicate students' CLA level.

TABLE VIII. RELIABILITY STATISTICS—PAIRED SAMPLES STATISTICS

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Teacher A	80.8529	34	3.7508	0.6433
Teacher B	80.2059	34	3.5486	0.6086

TABLE IX. PAIRED SAMPLES T-TEST OF TEACHER A AND B

	Paired Differences				**t**	**Df**	**Sig. (2-tailed)**
	M	***SD***	***95% Confidence Interval of the Difference***				
			Lower	**Upper**			
Pair 1 T A T B	0.647	2.043	-0.066	1.3599	1.84	33	0.074

This study proves that the test possesses external reliability. As shown in Table Ⅷ, the mean of the two sets of scores is 80.85 and 80.20, the standard deviation being 3.75 and 3.55. A parameter test, a paired sample *t*-test, was adopted to test the numerical values. As is shown in Table Ⅸ, the *t* value is 1.874, which is lower than the critical *t* value 2.045, and the significance is 0.74, which is higher than 0.05. Figures indicate that there is no significant difference between the two teachers' ratings. Therefore, the external reliability of the test has been fully obtained.

V. CONCLUSIONS

This paper explores the validity and reliability of a SET within the framework of CLT. The study leads to several conclusions as follows:

- The CALL environment supported the whole teaching process and the conduction of the tests. It functioned as an intermediate medium to build up a communication platform for the instructor and students, making it possible for each one to get involved in various speaking activities and interactions. Moreover, it helped the instructor collect students' recordings for the purpose of assessment and for use in further related research.

- Since all the speaking tasks in the post-test involved were similar to those in regular classes, students were in a comparatively relaxed manner taking the test. The scores, on the one hand reflected learners' inner language ability, and on the other hand, served as a method to evaluate the effect of the teaching method. This specific SET format had a positive backwash towards the teaching and learning in the EAVSC of the study.

- The design and conduction of the test adhered to the principle of being authentic or nearly authentic, which was the core idea of CLT model. The tasks relied essentially on students' real-life experiences, which ensured the realization of the face validity.

- By controlling such factors as test environment, individual factors and test factors, the authors provide prerequisites for gaining reliability. Through a series of calculation, including Cronbach's alpha coefficient, the Pearson correlation coefficient, descriptive statistics and paired t-test, it can be concluded that this communicative SET format achieved both internal and external reliability.

- The test proved to be valid in terms of face, content, construct and criterion validity through analyzing randomly selected subjects' recordings, comparing the test specifications and test content, computing the correlation coefficients between sub-tests and the whole tests and examining the degree of agreement between students' pre-test and post-SET scores. Data on validity and reliability show that the test reaches its objectives of analyzing students' oral performance.

Despite its positive contributions, the study has the following limitations:

- In this research, most of the subjects were with a CET-4 score over 545. Further generalizations to a larger sample or to other teaching and learning situations could not be deduced sufficiently.

- Though the test environments and test forms were similar enough to make students feel at ease, it is inevitable that factors such as anxiety might still exert a negative impact on some students so that the result from their post-SET scores, may not effectively reveal their real communicative competence.

REFERENCES

[1] P. Skehan, Progress in language testing: the 1990's. In Alderson and North (ed). Language Testing in the 1990s: The Communicative Legacy. London: Modern English Publications and the British Council, 1991, pp. 3-20.

[2] Lyle F. Bachman, Fundamental considerations in language testing. Oxford: Oxford University Press, 1990.

[3] Qiufang Wen, Testing and teaching spoken English. Shanghai: Shanghai Foreign Language Education Press, 1999.

[4] Ken Beatty, "Computer-assisted language learning," in Nunan, David. Practical English Language Teaching. Beijing: Higher Education Press, 2004.

[5] M. Warschauer and D. Healey, “Computers and language learning: an overview,” Language Teaching, 1998, vol. 31, pp. 57-71

[6] S. Bax, “CALL—past, present and future,” System, 2003, vol. 31, pp. 13-28.

[7] Baocheng Han, On Lyle F. Bachman’s communicative language testing model. Foreign Language Teaching and Research, 1995, pp. 55-60.

[8] Shen Zou, “On authenticity in oral testing,” Foreign Language World, 2001, vol. 3, pp. 74-78.

[9] Jigang Cai, “On the validity, reliability and practicability of compute-based CET-SET,” Foreign Language World, 2005, vol. 4, pp 66-75.

[10] Yuping Li, “An empirical study of the effect of the large-scale computer-assisted spoken English test,” Foreign Language World, 2009, vol. 4, 69-76.

[11] A. D. Cohen, Assessing language ability in the classroom. Beijing: Foreign Language Teaching and Research Press, 2005.

[12] A. Hughes, Testing For Language Teachers. Cambridge: Cambridge University Press, 1989.

[13] N. J. Anderson, “Individual differences in strategy use in second language reading and testing,” Modern Language Journal, 1991, vol. 75, pp. 460-472.

[14] J. C. Alderson, C. Clapham and D. Wall, Language test construction and evaluation. Cambridge: Cambridge University Press, 1995.

[15] J. M. Bland and D. G. Altman, Statistics notes: Cronbach’s alpha. British Medical Journal, 1997, pp.