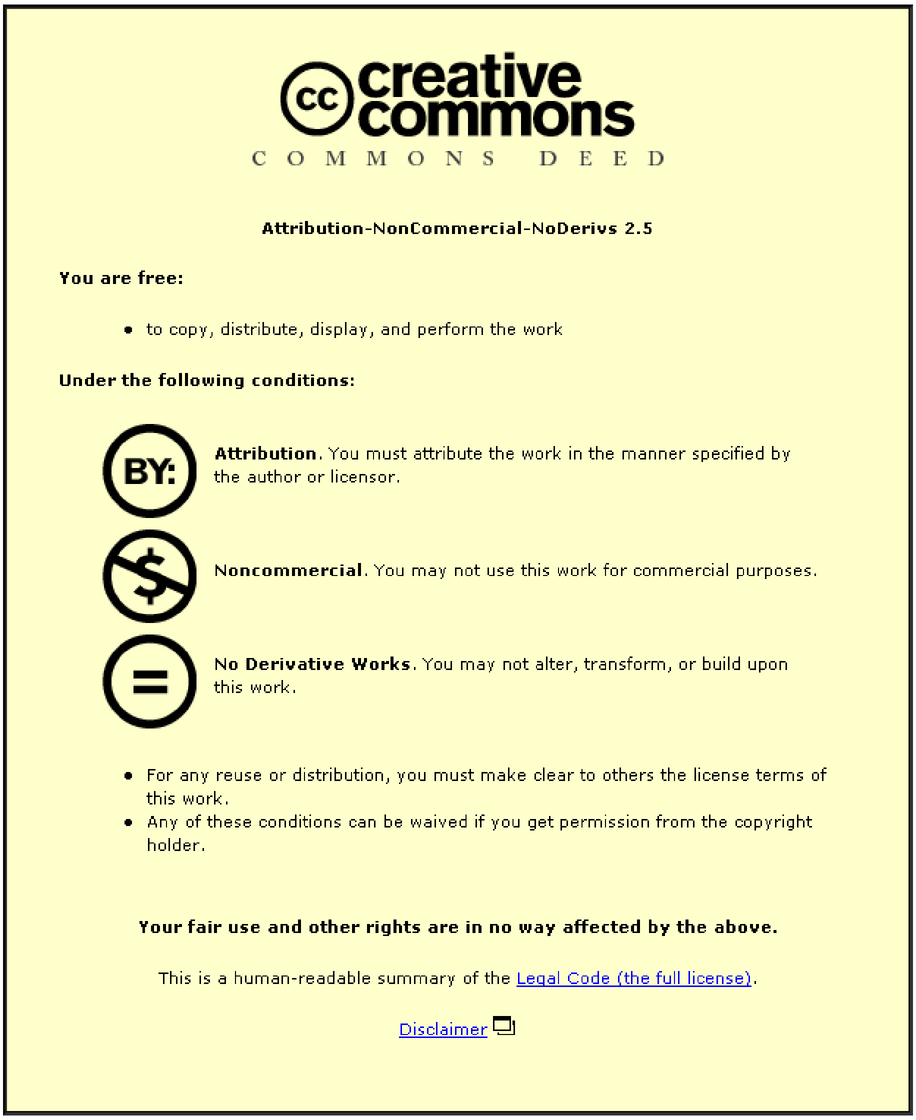

COMPUTER ASSISTED TESTING OF SPOKEN ENGLISH: A STUDY TO EVALUATE THE SFLEP COLLEGE ENGLISH ORAL TEST IN CHINA

Xin Yu and John Lowe

Computer Assisted Testing of Spoken English: A Study to Evaluate the SFLEP College English Oral Test in China

Xin Yu and John Lowe
University of Bath

Introduction

‘If you want to encourage oral ability, then test oral ability’ (Hughes, 1989:44)

Since its opening up to the outside world in the 1980s and the introduction of economic reforms that have involved engagement with the global economy and wider community, the Chinese government has become determined to promote the teaching and learning of English as a foreign language among its citizens. In particular, it has mandated the study of English for all college and university students and has made the passing of the College English Test (CET) at Band 4 level a requirement for obtaining a degree. With some ten million candidates annually (and rising) CET Band 4 has become the world’s largest language test administered nationwide (Jin and Yang, 2006). In a deliberate attempt to harness the backwash effect of examinations on teaching and learning, the Ministry of Education has insisted that all college and university students (generally when in their second year of study) must sit the CET Band 4 written papers that test reading, writing and listening skills in English. Aimed largely, but not exclusively, at those students majoring in English, there is also a higher level, Band 6, CET available.

A problem arises, however, when it comes to the formal testing of spoken English. There is a CET Spoken English Test (CET-SET), in use since 1999, which uses the widely accepted format for such assessments of a face-to-face interview with an examiner, together with a discussion on a given topic with two or three other students taking the test at the same time. This approach is both labour- and time-intensive, however, demanding highly skilled examiners as interlocutors and ‘small batch’ examining of students in sequence. As a consequence, simply for practical reasons of manageability of the test, CET-SET is only available to those who score higher than 80% in the Band 4 written tests or 75% in the Band 6 tests. Slightly conflicting data are available on the numbers taking this speaking test, with the lowest figure seen being around 40 000 and the highest around 90 000 (Jin and Yang, 2006: 22 & 30; Yang, personal communication, 2006). Whatever the precise figure, these data do indicate that over 99% of those students taking CET Band 4 written papers are not taking a test of spoken English. Even for those who do take the CET-SET, the stakes are not so high, since passing this test is not mandatory for obtaining a degree, unlike the CET written papers. The backwash implications of this are clear: neither among students learning English nor among teachers teaching it in China’s colleges and universities is there an emphasis on the development of spoken English proficiency. The high-stakes nature of CET Band 4 means that reading, writing and listening skills are taken seriously, but speaking skills receive much less attention, if any.

With this situation in mind, and with the recognition that computers are now a common feature of the higher educational environment in China, the Shanghai Foreign Language Education Press (SFLEP) and the University of Science and Technology of China (USTC), in Hefei, have been developing, since 2004, a computer-assisted speaking test, the SFLEP College English Oral Test System. (This produces the hideous acronym, SFLEPCEOTS, which will be substituted by CEOTS for the rest of this paper!) This test system is out of necessity, given current limitations of speech recognition software, something of a half-way house towards a fully computer-based assessment of speaking. It removes the need for a skilled examiner to be present during the conduct of the test but still requires such an examiner for the grading of the students’ performances. Students sit at a computer, log into the test system after a security test, and then respond to instructions on the screen. The test itself provides a variety of situations to which students respond in spoken English. Examples and details of the test items will be given during the presentation of this paper but not here. They include, however, responses to text, pictures and video clips, and even a discussion with two or three other students, randomly linked together. The students’ responses are recorded and then analysed and graded by examiners when they log into the system later. USTC use of this system over the last two years has shown that over 1500 students can take the test and have their performances graded in two or three days. It is argued, therefore, that CEOTS may present a more efficient system than the traditional face-to-face oral assessment and make regular testing of speaking proficiency on a large scale possible, while meeting the universities’ daily teaching needs in terms of its usability.

The University of Science and Technology of China has carried out some evaluation of the testing process, in terms of students’ perceptions and also of inter-marker reliability. This paper reports on a proposed joint study by USTC, SFLEP and the University of Bath, that will engage in a more thorough and wide-reaching evaluation of various aspects of this system and the possibility that it may offer an alternative to the current CET-SET that will open up the testing of speaking competence to the majority rather than a tiny minority of college students. This study is in its early phases and this paper will discuss some underlying conceptual issues and outline an evaluation research agenda, but will not report any of the provisional pilot data that have been collected so far. Although a considerable amount of data has been collected by USTC through the use of the test over two years, these data were not collected with a systematic evaluation of key aspects of the system in mind and it is recognised that further systematic data collection of various sorts is required. We would like to recognise at this point the generous support that has been afforded to the two presenters by professors Wu Min and Li Mengtao at USTC, who have been largely responsible for the development of CEOTS; but also to other colleagues at SFLEP and the National College

English Testing Committee who have offered and provided support to the development and implementation of the study.

Aims and Objectives of the Study

The research project that we are developing with our partners aims to evaluate the SFLEP College English Oral Test System with a specific concern over its potential for use as part of the CET programme in Chinese universities. It is recognised, however, that our findings may also have more generic implications for the use of computer-assisted English speaking tests, particularly with regard to the promotion of spoken English in Chinese universities. In order to gain acceptance for the test's large-scale public use we must establish its reliability and consider issues of efficiency and test manageability. If the test is to be acceptable as a replacement for traditional face-to-face oral English testing, however, the central concern is its comparative validity, and it is with issues of validity that this paper will primarily deal. Achievement of the additional aspiration to promote a positive backwash effect on English language teaching and learning by encouraging serious attention to be paid to spoken English depends not directly on the nature of the test itself but on whether it is adopted for high stakes use. This will depend on whether it becomes included with the tests of other language skills that must be passed in order to graduate. While willingness for such inclusion by the authorities – notably the CET board - will certainly depend on establishing the test's reliability and manageability, we argue below that at heart this remains a validity issue, drawing on Messick's concept of 'consequential validity' and the social impact of testing.

Specifically, the research questions that we have initially identified as guiding the evaluation of CEOTS are:

- How do the reliability and validity of the SFLEP College English Oral Test System compare with methods of face-to-face testing?
- Is the system efficient and manageable for use with very large numbers of students?
- What are the perceptions among users – both teachers and students – of the impact on English language teaching and learning of the introduction of this system?

The use of computer assisted tests of speaking proficiency is a relatively new field and, as yet, no large and detailed studies have been carried out to investigate the issues that we have identified, especially in relation to the situation in China. The first of the three questions above raises some broad issues that bring together three distinct fields: assessment, linguistic analysis and human-computer interaction. Before discussing a possible research agenda and methodological approach to address the research questions, it is important to identify concepts and theoretical approaches within these fields that we feel are particularly important.

Assessment: validity as a central concern

As Bachman and Palmer (1996) point out, the ideal outcome to any assessment regime is to achieve a balance among the essential qualities of validity, reliability, impact, and practicality to meet the requirements of the testing context. These qualities – or variations on them (e.g. Gipps 1994) – might usefully be taken to be the components of an evaluation of the assessment's 'fitness for purpose. Wolf (1998) identifies validity as being widely treated as the most crucial consideration in assessment. We concur and feel that - while recognising the importance of other concerns, particularly in a high-stakes context – validity remains the most significant issue in the context of CEOTS and its use in China. Our case depends, however, on a careful and contextualised interpretation of the concept of validity.

Traditional conceptualizations of test validity derived from psychometric testing (e.g. APA, AERA & NCME, 1966) treated validity in terms of three distinct facets, or evidential areas: construct validity, criterion validity and content validity. But, according to Messick, such a view is inadequate:

> *'Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment'(Messick, 1989:13).*

Validity as it is more widely understood today is an argument justifying certain interpretations to be drawn from or actions to be based on test scores; it is not actually the test that is valid, but rather the interpretations, conclusions and actions based on the test scores (Roever & McNamara, 2006).The crucial issues of test validity are 'the interpretability relevance, and utility of scores, the important or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use' (Messick, 1989:13). Black (1998) highlights the fact that all assessment processes are fundamentally social in character. Therefore test users should consider questions not only of how accurate a measurement is but also questions such as 'how valid are the interpretations made from the test data?' and 'how valid are the tests in terms of the decisions that are to be made?' As Kyriakides (2004) points out, validation should deal with issues concerning the consequences of test use (Kyriakides, 2004).

Messick's notion of consequential validity is central to making our case for the need to consider alternatives to face-to-face testing of spoken English in the Chinese CET context, in order to be able to assess the oral English competence of all students and not just a tiny proportion. It brings the backwash effect of the current CET-SET arrangements within the validity argument that has frequently been used to justify face-to-face testing as the 'most valid' form of assessment of speaking proficiency. This argument is generally based on an understanding of the construct validity of speaking tests that construes them as being more 'authentic' in their representation of the 'real-life' use of spoken language. We shall look at the construct validity of such tests in a moment but it is worth noting O'Loughlin's (2002) point that a language test need not reflect all aspects of ‘real-life’ communication (including gendered difference) in order still to be valid.

Tests are valid only for specific purposes (Madaus & Pullin, 1991) and in view of the complexity of the validation process, the suggestion that a test’s use or purpose should serve as a guide to validation is accepted (Worthen, Borg & White, 1993; Read & Chapelle, 2001). A difficulty clearly emerges when a test is intended to serve more than one purpose. In such a case it is likely that some form of compromise towards ‘optimal validity’ in relation to all the intended purposes must be sought. In the case of the CET system in China, two purposes may be identified. On the one hand, the test is intended to certify the individual’s competence as a speaker of English; on the other hand the test is intended as part of a strategy to promote more effective teaching and learning of English communication in all forms: reading, writing, listening and speaking. Clearly the two are not entirely separable in that if the speaking test does not validly certify spoken English competence but some other skill, it cannot serve to promote the teaching and learning of that competence. Thus, the argument goes, construct validity is the prime form of validity with which we must be concerned; although in the Chinese context outlined here, some way of dealing with the consequential validity issue of backwash also clearly needs to be sought. An important starting point in attempting to deal with this apparent tension in the purposes and their implications for validity is, therefore, to clarify how ‘construct validity’ might be understood in relation to the testing of spoken language.

Linguistic analysis: communicative competence

A speaking test can be defined as ‘a test in which a person is encouraged to speak, and is then assessed on the basis of that speech’ (Underhill, 1987:1). This minimal definition does not contain the point that any act of speaking serves a purpose, that purpose being communication. The ‘communicative competence’ approach to the teaching of language widely predominates in current practice; the replacement of more traditional, grammar and textual analysis models of language teaching by one which focus on developing communicative competence has been a major recent development in language classrooms in China. In relation, therefore, to a context of learning and teaching English, it seems reasonable that our interpretation of construct validity should be based on communicative competence models of language use and learning. Indeed, Heaton (1988) argues that construct validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills. The case is even stronger if we adopt Caroline Gipps’s (1994) suggestion that in an educational assessment regime ‘curriculum fidelity’ (where curriculum is to be interpreted broadly and not just in terms of subject content) is a more useful concept than that of construct validity that arose from the psychometrics testing tradition.

This is not the place to present a full account of communicative competence and its meaning for language teaching and learning, but the table below provides a useful summary of key aspects of the approach and a framework within which we may start to consider the construct validity of any form of speaking test that is intended to serve a communicative competence based curriculum.

Table 1: The components of a communicative competence model of language

Grammatical competence	Mastery of the language code
Sociolinguistic competence	Knowledge of appropriate language use
Discourse competence	Knowledge of how to connect utterances in a text so it is both cohesive and coherent
Strategic competence	Mastery of the strategies that speakers use to compensate for breakdowns in communications as well as the strategies they use to enhance the effectiveness of the communications

(Based on Canale and Swain 1980)

In principle at least, any given form of language assessment – whether of speaking or another skill – can be examined in relation to the extent to which it provides the opportunity for the candidate to demonstrate each of the above competences. But we should also be prepared to accept that no one form of assessment will assess all of the components equally well. As with any form of assessment, some sort of sampling of the domain will have to take place. This is revealed after a moment's thought about 'sociolinguistic competence', for example: clearly, even just within the domain of spoken language, the range of 'appropriate' language uses is enormous and way beyond the capacity of any manageable single assessment instrument to do anything more than lightly sample. In our comparisons of face-to-face and computer assisted modes of assessment we shall adopt the principle of asking what it is that each assesses, from the communicative competence model, rather than prioritising any component of that model in advance, thereby fitting an approach to validity that asks what interpretations can be made of performance in the assessment tasks.

Obviously, in the SFLEP College English Oral Test System, using computers instead of the interlocutors of a face-to-face test changes the participants in the speaking course. One of the participants in the interaction is changed to a computer, which may have potential effects on the students' language output. Almost no comparable research has been done between face-to-face and computer-assisted speaking tests, while there is some literature comparing tape-recorded tests with face-to-face speaking tests. The availability of visual as well as auditory stimuli is a key difference between computer based tests and those based on a recorded voice alone.

Kraut et al (1990) suggest that the visual channel is necessary to initiate a conversation in informal communications. As people talk, they are seeking positive understanding, such as acknowledgements, which take the form of gestures such as head nodding (Goodwin, 1981). Modes of body language, such as head gestures and facial expressions are well known to have strong effects on interactions in social situations generally (e.g. Argyle, 1983). It is therefore possible that the visual channel can affect the actual assessment of students' answers in an oral testing situation. For example, gestures of the hand could amplify a spoken explanation to advantage; whereas frowns could predispose the assessor in an unfavourable way towards a student (Seddon and Pedrosa, 1990). This has given support to the claim of Stansfield and Kenyon(1992) that the tape-recorded speaking test, in which there is no interlocutor , is 'fairer' than face-to-face speaking tests. Is the computer assisted test fairer than the face-to-face one? Savignon suggests communicative competence 'depends on the co-operation of all the participants involved' (1983:9). And part of the communicative competence is in knowing how to keep the conversation going, which includes knowing when to feign understanding and when to change the subject (Gunn, 2003). With no interlocutor involved in the computer assisted test, the issue of fairness and the capacity of items to test aspects of communicative competence will be will be important targets for data collection and analysis in this research.

The tape-based testing only covers some aspects of interactive speaking and the construct is more clearly connected with spoken production (Luoma, 2004). As with the tape-recorded test, the computer assisted test assesses only the spoken production of the testee rather than the interaction between testee and interlocutor found in interviews, role plays and other tests of speaking involving multiple speakers. The advantage of a computer assisted test is that the aural and visual stimuli remain precisely the same for all testees and, given the impersonality of the test procedure, differences due to inter-personal factors will be minimized. A question may occur in the test as to whether the response to such inauthentic stimuli can be regarded as authentic speech. Some believe that a face-to-face interview is most authentic because it is interactive. Underhill (1987) thought the voice-recorded test was not very authentic because the assessor of a recorded test can hear everything a live assessor can, but she cannot see the test, she therefore misses all the visual aspects of communication such as gesture and facial expression. A crucial question defining authenticity is 'authentic to what?' (Messick, 1994:18). Authenticity is not an objective quality as such; it is subjective and dependent on who is judging the authenticity (Gulikers, et al, 2006). There is almost no literature about authenticity of computer assisted speaking test; therefore in the research this issue will be investigated in detail.

Being afraid of poor performance in front of other people, students tend to be silent in class. This is particularly noticeable in Asian English as Second Language learning classes. In the 1990s related studies indicated that students who used to be shy in face-to-face discussion and who were considered low achievers in language learning became more active participants in computer-assisted classroom discussion (Beauvois, 1992, 1995; Kelm, 1992). Without seeing each other in the test, with a less threatening means to communicate, students may find it easier to speak. In the computer-assisted tests, will the testees find it easier to speak, in the absence of visible testers? Will their language output be changed?

The essential challenge from advocates of face-to-face testing is that computer assisted assessment is not an authentic simulation of 'real life' language use. We would counter that face-to-face exchanges actually only represent on spoken language context. For students who may go on to an academic career, for example, we suggest that the ability to make a presentation – or give a lecture, if you will – to a large audience on a familiar topic may be a skill that will be required in the future. Furthermore, in this era of electronic communications, speaking over the telephone, or via an internet link, such as Skype, with or without visual contact, is something that will be a common part of these students' lives – perhaps more common than face-to-face encounters for many. But the argument is not so much one of which of these assessment contexts is 'more authentic' but rather that we should ask what forms of spoken language use any assessment best approximates to and may therefore for which it may claim some level of validity.

Human-computer interaction: who are you talking to?

With rapid developments in computer-based technologies in recent years, the use of computers to administer tests is becoming increasingly common in education (Bonham et al, 2000; Mason et al, 2001; Olson, 2002). It is predicted that the use of computer-assisted tests for language assessment and other assessment purposes will become increasingly predominant in the immediate future (Bennett, 1999). However some researchers argue that these and other computer-linked factors may change the nature of a task so dramatically that one cannot say the computer-assisted and conventional version of a test are measuring the same thing (McKee& Levinson, 1990).

Changing the administration of the test may affect the reliability of a test. Computer-based test provides testees with an equal opportunity by allowing every testee to have the very same testing experience. Introducing a new method of assessment however may cause students anxiety. For many people, the test situation itself creates considerable anxiety which can badly affect their performance (Underhill, 1987). However, Foot (1999) highlights that students may not necessarily perform better if they are more relaxed. In general, higher-attaining students will adapt most quickly to any new assessment approach (Watson, 2001; Noyes, 2004) and will quickly develop test-taking strategies that benefit them in the new approach. Computer anxiety is another potential disadvantage that may affect test performance (Henning, 1991). Also differences in the degree to which students are familiar with using computers may lead to differences in their performances on computer-assisted or computer-adaptive tests (Hicks, 1989; Henning, 1991). Clark (1988) and Stansfield et al (1990) found that examinees sometimes felt nervous taking a computer assisted test, because of a feeling of lack of control. Some examinees reported that they felt this nervousness prevented them from doing their best.

Research into computer-supported learning suggests that women suffer from lower levels of computer literacy and lower confidence levels in its use (Yates, 2001). Men and women were also observed to behave differently in on-line group discussion (Barrett & Lally, 1999). In particular, it was observed that men's talk was, typically, more numerous and longer than that of women, and tended to include greater levels of social exchange. Women, however, appeared, typically, to be more interactive than men. Some studies claim that the internet increases engagement, confidence, and responsibility with a less threatening means to communicate (Chun, 1994; Beauvois, 1995; Skinner & Austin, 1999), while McGrath (1997-98) found that those students who do well in a face-to-face environment may be suppressed in a web-based environment and vice versa. There is a large body of research in the field of gender, familiarity and anxiety on human & computer interaction, while almost no research has been done on comparing differences in behaviour and speech when human beings are speaking to a computer rather than to other human beings.

Towards a Research Agenda

It is clear from the discussion above that a full evaluation of CEOTS, even just in relation to its possible use in the CET system, demands the investigation of many factors. We further recognise, however, that given the very recent appearance of computer assisted assessment of spoken language and the apparent shortage of research into speech based interactions with computers this is an opportunity to carry out more fundamental research that goes beyond an evaluation of a particular system.

As suggested above, we believe that the starting point for our evaluative research is to ask what interpretations can be validly made of performance in any given form of spoken language assessment, rather than to start with a notion of what is 'authentic' or 'non-authentic'. This suggests to us that one of our chief research tasks is to analyse the spoken language generated under various testing circumstances and by different tasks set within those circumstances. We are fortunate in having a huge volume of test results – including the actual voice recordings – available to us through our collaboration with USTC. We also have the interest and co-operation of the CET administration in this project and through them will have access to video recordings of a large number of their face-to-face tests. These clearly present opportunities for detailed linguistic analysis of the responses generated by different item formats and individual items in the test, for some of which we shall use linguistic analysis computer software. The precise nature of the aspects of language we shall be looking for remain to be firmly established but we hope that we shall be able to produce a 'profile' of language responses to testing modes and item types.

Despite the existence of this considerable database, however, we feel there is a need to collect data under more controlled conditions. We are in particular interested in investigating testee responses beyond the linguistic and plan to video individuals taking the computer assisted tests to allow us to analyse face and body activity. We would also like to examine their subjective perceptions of the two forms of testing, which we shall do through both quantitative survey techniques and suing individual and group interviews to obtain data for qualitative analysis.

Interviews will also be held with English teachers at USTC, particularly those who have experience of teaching before and after the introduction of CEOTS, to obtain –admittedly somewhat subjective – data on the impact that the testing has had on their language classes.

Data on aspects of the reliability of the speaking test results will be collected in a variety of ways. Test-retest reliability data will be generated for selected groups of students. The grading process will be subject to scrutiny by observation of the process, through interviews with markers and by comparing marks from different markers for the same recordings. Some similar data will be collected for face-to-face tests, and it is hoped that the co-operation that the national CET committee have promised us will give us access to their own data on the reliability of the CET-SET.

Finally, issues of manageability of the system, particularly with respect to a potential huge increase in scale, will be examined through discussions with USTC staff involved in the system management and development.

Conclusions

Computer assisted speaking testing is a relatively new field and there are, as yet, few large and detailed studies in this field. Using computers can potentially allow simultaneous performance of the speech production part of testing by a large number of students, although the grading of their performances remains labour-intensive. Whether a system such as CEOTS can overcome the inefficiency of traditional face-to-face testing and make oral testing on a large scale possible, without major detrimental impact on the validity and other aspects of the assessment, is at the heart of this research. We recognise the complexity of the project that we are taking on and anticipate that we shall be continually reviewing both our methodological and theoretical approaches. We remain convinced, however, that alternatives to face-to-face testing must be found so that the annual ten million plus English language testing candidates in Chinese universities can be offered a test of their speaking competence. If this is not done, then the backwash effect of the absence of such an examination for the vast majority will continue to distort the teaching and learning of English among those students and to undermine the government's attempts to improve the language proficiency of the country's university graduates. The research in which we are currently engaging promises therefore to be of considerable practical and theoretical significance.

References

American Psychological Association (APA), American Educational Research Association (AERA) and National Council on Measurement in Education (NCME) (1966) *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Argyle, M. (1983) *The Psychology of Interpersonal Behavior* (4th ed). Harmondsworth: Penguin.

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Barrett, E. and Lally, V. (1999) Gender differences in an on-line learning environment. In *Journal of Computer Assisted Learning*, Vol. 15, pp48-60.

Beauvois, M. H. (1992) Computer-assisted classroom discussion in the foreign language classroom: Conversation in slow motion. In *Foreign Language Annals*, Vol. 25(5), pp 455-464.

Beauvois, M. H. (1995) E-talk attitudes and motivation in computer-assisted classroom discussion. In *Computer and he Humanities*, Vol. 28, pp177-190.

Bennett, R. E. (1999) Using new technology to improve assessment, In *Educational measurement: Issues and practice*, Vol. 18(3), pp5-12.

Black, P. (1998) Assessment and classroom learning. In *Assessment in Education*, Vol. 5, pp7-74.

Bonham, S. W., Beichner, R. J., Titus, A., and Martin, L. (2000) Education research using Web-based assessment systems. In *Jonrnal of Research on Computer in Education*, Vol. 33, pp28-45.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. In *Applied Linguistics,* Vol. 1(1), pp8-24.

Chun, D. (1994) Using computer networking to facilitate the acquisition of interactive competence. In *System*, Vol. 22, pp17-31.

Clark, J. L. D. (1988). Validation of a tape-assisted ACTFL/ILR-scale based test of Chinese speaking proficiency. In *Language Testing*, Vol. 5(2), pp197-205.

Foot, M. C. (1999) Relaxing in Pairs. In *ELT Journal,* Vol. 53(1), pp36–41.

Gipps, V. C. (1994) *Beyond testing-towards a theory of educational assessment.* Washington, D.C.: The Falmer Press.

Goodwin, C. (1981) *Conversational organization: interaction between speakers and hearers*. New York: Academic Press.

Gulikers, J., Bastiaens, T. and Kirschner, P. (2006) Authentic assessment, student and teacher perceptions: the practical value of the five-dimensional framework. In *Journal of Vocational Education and Training,* Vol. 58(3), pp337-357.

Gunn, L. C. (2003) Exploring second language communicative competence. In *Language Teaching Research*, Vol. 7(2), pp240-258.

Heaton, J. (1988) *Writing English Language Tests*. London: Longman

Henning, G. (1991) Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (New York: Newbury House), pp209-222.

Hicks, M. (1989) The TOEFL computerized placement test: Adaptive conventional measurement. *TOEFL Research Report No. 31*. Princeton, NJ: Educational Testing Service.

Hughes, A. (1989) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Jin, Y. and Yang, H. Z. (2006) The English proficiency of college and university students in China: as reflected in the CET. In *Language, Culture and Curriculum*, Vol. 19(1), pp21-36.

Kelm, O. R. (1992) The use of synchronous networks in second language instruction: a preliminary report. In *Foreign Language Annals*, Vol. 25, pp441-545.

Kraut, R.E., Fish, R.S., Root, R.W. and Chalfonte, B.L. (1990) Informal Communication in Organizations: Form, Functions and Technology. In: Oskamp, S. and Spacapan, S. (Eds.) *Human Reactions to Technology. The Claremont Symposium on Applied Social Psychology.* Beverly Hills, CA: Sage Publication.

Kyriakides, L. (2004) Investigating validity from teachers' perspectives through their engagement in large-scale assessment: the emergent literacy baseline assessment project. In *Assessment in Education*, Vol. 11(2).

Luoma, S. (2004) *Assessing speaking*. Cambridge: Cambridge University

Press

Madaus, G. and Pullin, D. (1991) To audit and validate `high stakes' testing programs, in: R. G.. O'Sullivan (Ed.) *Advances in program evaluation*: Vol. 1A, Effects of mandated assessment on. teaching (Greenwich, CT, JAI Press), pp139-158.

Mason, B. J., Patry, M., and Bernstein, D.J. (2001) An examination of the equivalence between non-adaptive computer-based test and traditional testing. In *Journal of Educational Computing Research*, Vol. 24, pp29-39.

McGrath, C. (1997-98) A new voice on interchange: is it talking or writing? Implications for the teaching of literature. In *Journal of Educational Technology systems*, Vol. 26(4), pp291-297.

McKee, L. M. and Levinson, E. M. (1990) A review of the computerized version of the self-directed search. In *Career Development Quarterly*, Vol. 38(4), pp325-333.

Messick, S. (1989) Validity. in R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. New York: American Council on Education & Macmillan.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessment, in *Educational Research*, Vol. 23(2), pp13-23.

Noyes, J., Garland, K. and Robbins L. (2004) Paper-based versus computer-based assessment: is workload another test-mode effect? In *British Journal of Educational Technology*, Vol.35 (1), pp111-113.

O' Loughlin, K. (2002) The impact of gender in oral proficiency testing. In *Language Testing*, Vol. 19(2), pp169-192.

Olson, A. (2002) Technology solution for testing. In *School Administration*, Vol. 59, pp20-23.

Read, J. and Chapelle, C. A. (2001) A framework for second language vocabulary assessment. In *Language Testing*, Vol. 18(1), pp1-32.

Roever, C. and McNamara, T. (2006) Language testing: the social dimension. In *International Journal of Applied Linguistics*, Vol. 16(2).

Savignon, S. (1983) *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley

Seddon, G. M. and Pedrosa, M. A. (1990) Non-Verbal Effects in Oral Testing. In *British Educational Research Journal*, Vol. 16(3), pp305-310.

Skinner, B. and Austin, R. (1999) Computer conferencing: Does it motivate EFL students? In *ELT Journal*, Vol. 52(1), pp38-42.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle,y F., Ulsh, I., and Cowles, M. A. (1990). The development and validation of the Portuguese Speaking Test. In *Hispania*, Vol. 72, pp641-651.

Stansfield, C.W., & Kenyon, D.M. (1992) The development and validation of a simulated oral proficiency interview. In *Modern Language Journal*, Vol. 76, pp 129-141.

Underhill, N. (1987) *Testing spoken language*. Cambridge: Cambridge University Press.

Waston, B. (2001) Key factors affecting conceptual gains from CAL. In *British Journal of Educational Technology*, Vol. 32(5), pp587-593.

Wolf, R. M. (1998) Validity issues in international assessments. In *International Journal of Educational Research,* Vol. 29(5), pp491-501.

Worthen, B. R., Borg, W. R. and White, K. R. (1993) *Measurement & evaluation in the schools.* London: Longman.

Yates, S. J. (2001) Gender, Language and CMC for education. In *learning and Instruction*, Vol. 11, pp21-34.