# Computer Assisted Testing of Spoken English:

# A Study of the SFLEP College English Oral Test System in China

**John LOWE & Xin YU**
**Department of Education, University of Bath,**
**Bath, BA2 7AY, United Kingdom**

## ABSTRACT

This paper reports on the on-going evaluation of a computer-assisted system (CEOTS) for the assessing of spoken English skills among Chinese university students. This system is being developed to deal with the negative backwash effects of the present system of assessment of speaking skills which is only available to a tiny minority. We present data from a survey of students at the developing institution (USTC), with follow-up interviews and further interviews with English language teachers, to gauge the reactions to the test and its impact on language learning. We identify the key issue as being one of validity, with a tension existing between construct and consequential validities of the existing system and of CEOTS. We argue that a computer-based system seems to offer the only solution to the negative backwash problem but the development of the technology required to meet current construct validity demands makes this a very long term prospect. We suggest that a compromise between the competing forms of validity must therefore be accepted, probably well before a computer-based system can deliver the level of interaction with the examinees that would emulate the present face-to-face mode.

## INTRODUCTION

Reflecting the Chinese governments' determination to promote the teaching and learning of English among its citizens, all Chinese university students must now pass the College English Test (CET) at Band 4 level (or an equivalent test) as part of their degree programme. With over ten million candidates annually (and rising) CET Band 4 has become the world's largest language test administered nationwide [1].

The English speaking component of this test (CET-SET) is, however, only available to a tiny minority of students, because of the adoption of a direct, face-to-face testing mode: over 99% of those taking Band 4 written papers are not taking a test of spoken English. The backwash implications of this are clear: neither among students learning English nor among teachers is there an emphasis on the development of spoken English proficiency. This conflicts with the College English Curriculum objective that students should develop an 'ability to use English in a well-rounded way, especially in listening and speaking.'

The most recent formulation of the College English Curriculum (2007) stresses the importance of computer- and web-based teaching models, especially for the training of speaking and listening abilities. With this situation in mind, the Shanghai Foreign Language Education Press (SFLEP) and the University of Science and Technology of China (USTC), in Hefei, have been developing a computer-assisted speaking test, the SFLEP College English Oral Test System (CEOTS). This test system is something of a half-way house towards a fully computer-based assessment of speaking, removing the need for a skilled examiner to be present during the conduct of the test but still requiring an examiner to grade students' performances. The test itself provides a variety of situations to which students respond in spoken English. These responses are recorded and then graded later by examiners when they log into the system. USTC use of this system has shown that over 1500 students can take the test on one site and have their performances graded in two or three days. It is argued, therefore, that CEOTS may present a more efficient system than face-to-face assessment and make regular testing of speaking proficiency on a large scale possible, while meeting the universities' needs in terms of its usability.

## THE STUDY

This paper reports on a joint study by USTC, SFLEP and the University of Bath, to carry out a 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 of students. This study is on-going and this paper reports on the issues, aims and approaches to the evaluation, and some initial results.

We recognise 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. Through a comparison with face-to-face tests, our study investigates the reliability, validity, efficiency, management, social consequences and backwash on teaching and learning of this computer-assisted speaking test. We address the following questions:

- How do the reliability and validity of CEOTS compare with those 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?

These questions bring together three distinct fields: assessment, linguistic analysis and human-computer interaction. Before discussing our methodology, we identify concepts and theoretical approaches within these fields that we feel are particularly important and inform our analysis of the collected data.

## THEORETICAL CONCERNS

### Assessment: validity as a central concern

As Bachman and Palmer [2] point out, the ideal outcome to any assessment regime is to achieve a balance among validity, reliability, impact, and practicality to meet the requirements of the testing context. These qualities – or variations on them [3] – might usefully be taken to be the components of an evaluation of the regime's 'fitness for purpose'. Two purposes of the CET system may be identified: certifying the individual's competence as a speaker of English; and supporting the promotion of more effective teaching and learning of English.

We concur with Wolf [4] that validity is the most crucial consideration 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. Our case depends, however, on a careful and contextualised interpretation of the concept of validity.

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 must also be sought. Messick's notion of consequential validity [5] is central to making our case for the need to consider alternatives to face-to-face testing in the CET context, in order to be able to assess the oral English competence of all students and not just a tiny proportion.

### Linguistic analysis: communicative competence

The 'communicative competence' approach to the teaching of language widely predominates in current practice. 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. In our comparisons of these two modes of assessment we 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.

Almost no comparative research has been done between face-to-face and computer-assisted speaking tests, although there is some literature comparing tape-recorded tests with face-to-face speaking tests. With no interlocutor involved in the computer assisted test, the issue of fairness and the capacity of items to test aspects of communicative competence are important targets for data collection and analysis.

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. 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 therefore for which it may claim some level of validity.

### Human-computer interaction: who are you talking to?

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 [6]. 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 [7].

Introducing a new method of assessment may cause students anxiety; computer anxiety is another potential disadvantage that may affect test performance [8]. Clark [9] and Stansfield et al [10], for example, found that examinees sometimes felt nervous taking a computer assisted test, because of a feeling of lack of control.

Research into computer-supported learning suggests that women suffer from lower levels of computer literacy and lower confidence levels in its use [11]. There is a large body of research in the field of gender, familiarity and anxiety on human-computer interaction, while almost no research can be found comparing differences in behaviour and speech when human beings are speaking to a computer rather than to other human beings.

## DATA SOURCES

A questionnaire was administered to students at USTC, who had some experience of CEOTS. This university enjoys quite a high reputation in China and is known for sending many students abroad – particularly to the USA - for further study after their first degree. It might be expected, therefore that the students here would rate the importance of English rather more highly than in many Chinese universities. The aim of the questionnaire was to produce a quantitative descriptive account of attitudes to and experiences of learning English and of both CEOTS and more conventional face-to-face testing of spoken English, and to look for patterns in terms of gender and course of study. A total of 660 valid questionnaires was returned.

Questionnaire responses were used to identify students for follow-up interviews, both individual and group-based. These interviews sought to explore in greater depth perceptions of and attitudes towards learning English and the different forms of spoken language testing. Interviews were also held with English teachers and other staff involved in CEOTS to obtain their views on the same issues and on the management of the test and, in particular, the grading process.

Collaboration with USTC gives us access to a huge volume of testing results – including the actual voice recordings. Some of these are being used for detailed linguistic analysis of the responses generated by different item formats and individual items in the test but only limited findings from this analysis are available at this time and will not be a primary focus in this paper.

Further test results were obtained under more controlled experimental conditions to generate voice recordings from the same students in both face-to-face and computer assisted tests. These tests were video-recorded and students were also interviewed immediately afterwards

## SOME INITIAL FINDINGS

Rather as expected, the questionnaire data revealed that the students held generally positive attitudes to learning all aspects of English, with 94% agreeing that learning English is important for their future. Although reading was seen as the most important skill in English, with 98% agreeing that it is important, speaking was not far behind at 96%. Interestingly, a large majority (88%) rejected the statement that English is only important for those who intend to travel abroad. When it came to their own performance in spoken English, however, only 13% expressed confidence in their ability.

Evidence of negative backwash from the testing regime is provided by the mere 21% who acknowledged practicing spoken English regularly and the 37% who stated that speaking practice was not a significant part of their English classes. 53% admitted that they would only practice English if it were necessary to pass a spoken English test. Although the CEOTS test is compulsory for all students in USTC, it contributes only 5% to their total English marks and students admitted in interviews that this tiny contribution led them largely to ignore it.

No statistically significant differences were noted between those majoring in different subjects, but a few differences between responses from male and female students were significant, though small. Females students tended to rate the importance of English slightly higher than males and tended to be more confident in their spoken English ability. Male students, on the other hand, expressed greater confidence in using a computer than females.

All of the respondents had experience of both CEOTS and face-to-face testing and only 13% expressed a preference for the former, with 63% preferring face-to-face testing. This preference for face-to-face testing was echoed in the opinion of 52% that they performed better in this format whereas only 16% felt that CEOTS elicited a better performance. It was difficult to identify clear reasons for this from the questionnaire data, although a clue appears in the fact that over half of the students (53%) agreed with the comparison between speaking to a computer and speaking to a wall! This compares with 71% who declared that they enjoyed interacting with a 'live' examiner.

The interviews confirmed and fleshed-out data from the questionnaire responses and allowed an exploration of the students' experiences with assessment. As in the sample as a whole, the majority of those interviewed preferred face-to-face testing and claimed it as a more authentic modelling of 'real-life' use of spoken language. This was always judged in terms of the extent to which one was responding to a live and present human being, and there was widespread recognition that communication through speaking was more than just 'using words'. The importance – the necessity – of body language from both the examiner and the examinee was stressed by many. For example, one student, who felt that face-to-face testing was 'like chatting' and 'chatting is the best way to practice spoken English' commented on the 'inspirational' value of 'gestures or facial expression' from the examiner 'to let you think on your own'. Another noted that a small gesture or even just eye contact was helpful. Although teachers that we spoke to declared that in face-to-face testing they tried to avoid facial expression or gesture, it was clear from student comments that they too found it impossible to maintain a live conversation without giving some, perhaps involuntary, responses, usually a nod or a smile. These not only assisted the examinees but gave the exercise greater authenticity in their eyes as it encouraged them to continue with their attempts to speak, in a way that they might expect from another speaker in a real-life context.

The students were, however, quite aware that a test is a test and all admitted to nervousness brought on by this recognition. Nervousness in a face-to-face context was, however, quite different from that experienced when faced with an impassive computer and a fixed time to respond, being inexorably ticked out by a time bar at the bottom of the screen. One interviewee described his nervousness as a sense of 'excitement', which was greater when facing people than facing a computer and that this excitement meant that '*I can start the test more easily and start to talk*'. Interestingly, nervousness in front of the computer was often expressed in terms of worrying about making mistakes. Because anything spoken was recorded, it was felt that mistakes could not be corrected and would be marked. This led to an artificial concern with accuracy rather than communication, with grammar and vocabulary rather than ideas and content of what was said. The lack of reaction from the computer and a fixed time given for a response led students to feel the need to fill the time with talking – with words – rather than being concerned with what they said. They missed the 'natural' termination signs or encouragement to continue that they expected with a live audience.

Students were aware that they would tend to produce a different type of language in the computer test, trying to focus on 'accuracy' or 'grammar'. Preliminary comparative analysis of recordings from both types of testing indicates that this may be true, although this analysis is in its early stages and we are still experimenting with various forms of linguistic analysis here. Language used in the face-to-face tests tends to include more pauses and to be repetitious, with less accurate grammar. Sentence construction and vocabulary are generally more diverse, complex and accurate in CEOTS performances. There is also evidence that in the preparation time for each item on CEOTS many students develop a list of vocabulary and phrases that they try to work into their responses, often trying simply to fill all the available response time with as much of these as they can. On the other hand, weaker students often latch on to the 'hints' that they are given for many items, either simply repeating these or re-ordering the same vocabulary and constructions in their responses.

One interviewee justified his attention to accuracy in CEOTS as he expected the grading of his performance to be more careful, more thorough, since the recording could be returned to and replayed by the examiner in an attempt to get 'the right mark', whereas the face-to-face interview was 'over when it was over' and marks would be given on impressions made by the examinee. This view of the grading process was shared by several other students and it was only on this issue of grading reliability – which they saw as test 'fairness' – that the students rated CEOTS above face-to-face assessment. The latter was assumed to be more open to examiner bias, particularly through the examiner's response to the student rather than the language: '*The examiner will mark according to his mood, your appearance when you speak, or many other things*'. This perceived objectivity in CEOTS grading must be set against the common view, however, that the passive nature of the test prevents one from producing one's best performance, so that the overall view was that face-to-face provides a better test of one's speaking ability, but the computer-based assessment is more accurately and fairly graded.

In fact, the teachers involved with the marking admitted that their marking of the CEOTS recordings was often rushed and impressionistic. Despite the capacity of a computer system to batch process large numbers of students simultaneously, the grading process remains sequential and examiners may be expected to grade large numbers of recordings in a short time. This is tedious and teachers admitted to making judgements of a student's overall performance based on just part of the recording – essentially impressionistic marking. There is no time in fact to mark on detailed points of grammar or accuracy and a general sense of 'fluency' was declared to be the main marking criterion. Perhaps surprisingly, given the demand on their time made by face-to-face testing, the English language teachers preferred this to the computer test. Their reasons were very similar to those of the students, centring on greater 'authenticity' that arises from the human interaction in the test, which they saw as essential to the very purpose of the use of spoken language. To put it briefly, the teachers held the computer based assessment to suffer from much lower construct validity, while not sharing the students' views about its greater comparative reliability.

Teachers did see value in CEOTS as a basis for formative assessment and an opportunity for the speaking practice that the students so clearly lack. They – and the students – made suggestions for improving the particular form of the test. These included broadening the range of item types but the main concern of both groups was to find ways of making the items more 'true to real life', through the inclusion of more video-clips, including a real person asking the questions rather than a dis-embodied voice. Beyond these suggestions for improving the content and presentation, however, there remained a firm belief that a computer based system could not provide a valid alternative to traditional modes.

## DISCUSSION AND CONCLUSIONS

In a discussion of these findings it is important to distinguish between computer based assessment of spoken language *per se* and the particular form it takes in CEOTS. It is also important that any evaluation of either be placed in the context within which the assessment is to operate and the purposes that it is expected to meet. In this case the context is one of a policy of promoting competence in English among a vast student population and the purposes of testing are both to certify that competence and to act as a stimulus to individuals to improve their competence – with that competence being defined across all four components of language use. There is evidence in our own findings, if indeed it were needed, that high-stakes assessment such as the College English Test does have a significant ‘backwash’ effect and that this contributes to the importance with which the Chinese students view the learning of English. The negative backwash effect of the absence of a speaking test component in CET is also clearly seen. It is noticeable, however, that this negative effect has not been effectively countered at USTC by the inclusion of a spoken English test for all, simply because the proportion of the final marks given for the test is so small as to indicate that the authorities do not take it seriously either.

At the national level, therefore, action to counter the negative backwash effect of CET on spoken English must comprise both a compulsory test of speaking skills for all students and that this test be given significant weighting in determining a student’s overall grade. The problem of meeting the first of these requirements was what opened this paper and what led to the development of CEOTS as a means of mass testing speaking skills. The second requirement can in practice only be met if an acceptable standard of reliability of the grading can be achieved. This was shown to be a problem with CEOTS as it is presently managed, largely because of the tedious nature of the process for the graders. In principle this does not differ from the process of grading essay type examination papers, however, and presumably moderation processes similar to those used in essay marking could be put in place.

In the long term, the ideal would be to have the grading also carried out by computer. This is already available in some publically available tests for a limited range of speaking skills and criteria for judging them. USTC and its partners are already experimenting with computer-based grading of the reading out loud of a set passage. It is recognised, however, that our present level of technology and software development do not allow reliable grading across the range of criteria that might be demanded for a valid test of the sort CET is intended to be, and which can cope with the range of speaking competencies that would be expected.

The other aspect of developing adequate levels of grading reliability is the clear specification of criteria for this grading. Criteria do exist for CET testing, at least in the form of the specification of domains to be considered (flexibility, appropriateness, coherence, accuracy, size and range) and the assessment of these is clearly beyond the present limits of our computer technology. But at the heart of this specification of assessment criteria is a more fundamental issue that we feel is of primary concern in making judgements about any computer based system and which is actually an issue of validity rather than reliability.

Both the traditional face-to-face test and CEOTS can claim face validity in that they are both tests of spoken English. To go any further with validity claims, however, requires a clear specification of the ‘construct’ of ‘spoken English’, or ‘spoken English competence’. Here we are up against problems of the sheer complexity of the construct, arising largely from the diversity of situations in which spoken English may be used. Not all of this involves face-to-face interaction of the form that traditional testing takes; one could think of telephone conversations, delivering a lecture or speech, or giving a radio commentary on an event as examples of ‘authentic’ use of speaking that are not the one-to-one model of traditional tests. These different speaking contexts commonly demand different speaking skills that may or may not be associated with other communication tactics such as the use of facial expression and body language. All are authentic and which associated set of skills is to be prioritised in teaching and assessing will presumably depend on the purpose or purposes for which the language is being learned. Unfortunately, the CET specification of the purposes for which spoken English is to be learned is rather broadly expressed. The responses of our students indicate that they themselves see spoken English skills as being important mostly for those who will eventually travel abroad and for these, the conversational skills modelled in face-to-face testing would be useful. But this applies to only a small minority, leaving us to ask in what sort of situations the majority might use spoken English and what English speaking skills they might require. This is at the heart of questions about the validity of any testing system that we use for the mass testing of spoken English skills.

At this point we can bring Messick’s notion of consequential validity [5] into the picture, arguing that the current situation involves a tension between construct and consequential validity of spoken English testing in China. The large negative backwash effect of the absence of mass testing now becomes part of the validity debate and if we ask about the validity of the current spoken English testing regime we must offset the claims of high construct validity made for face-to-face testing with their negative consequential validity. At present, the commitment is entirely to the former at the expense of the latter and we might ask whether a better balance can in fact be achieved; or, in the context of this paper, whether a computer-based testing system can bring about a better balance.

Clearly, present levels of computer technology do not allow the levels of interaction that characterise face-to-face testing or the conversational interaction that it emulates. To achieve this we are probably demanding levels of machine-based artificial intelligence that are being sought but far from being achieved, leaving our 'ideal test' on a distant horizon. Progress towards more sophisticated levels of human-computer interaction is being made and even the present level of developments here would allow CEOTS to be made more interactive (although probably with greater system demands that might be a problem for its more widespread use). The 'humanisation' of CEOTS or any other system, that our students expressed a desire for will be an incremental process and may reach a point at which the test is deemed to be 'acceptably valid'. Alternatively, we can indulge in some redefinition of the domain of spoken language use that would alter our validity demands, but this may meet opposition from those with a professional interest in maintaining current definitions that would make it harder to achieve than the developments in technology.

Ultimately, however, we would argue that if China is intent on promoting a broad range of English competencies among its university students it must take the negative consequential validity of its present system into account. A point will be reached at which the prize of removing this backwash effect will be worth the price that may be perceived to be demanding by accepting reduced (or altered) construct validity. CEOTS, as outlined and discussed in this paper, does not yet offer that position but we are confident that it is only through the development of computer-based systems of spoken language assessment that a solution to the dilemma will be found.

## REFERENCES

[1] Y. Jin and H.Z. Yang, "The English proficiency of college and university students in China: as reflected in the CET", **Language, Culture and Curriculum**, Vol. 19, No. 1, pp. 21-36, 2006.

[2] L. Bachman, L. and A. Palmer, **Language Testing in Practice**, Oxford: Oxford University Press, 1996

[3] Caroline Gipps, **Beyond Testing**, London: Falmer Press, 1994.

[4] R.M. Wolf, "Validity issues in international assessments", **International Journal of Educational Research**, Vol. 29, No. 5, pp. 491-501, 1998

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

[6] R.E. Bennett, "Using new technology to improve assessment", **Educational measurement: Issues and practice**, Vol. 18, No 3, pp. 5-12, 1999

[7] L.M. McKee and E.M. Levinson, "A review of the computerized version of the self-directed search", **Career Development Quarterly**, Vol. 38, No. 4, pp. 325-333, 1990.

[8] G. Henning, "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, pp. 209-222, 1991.

[9] J.L.D. Clark, "Validation of a tape-assisted ACTFL/ILR-scale based test of Chinese speaking proficiency", **Language Testing**, Vol. 5, No. 2, pp. 197-205, 1988.

[10] C.W. Stansfield, D.M. Kenyon, R. Paiva, F. Doyle, I. Ulsh and M.A. Cowles, "The development and validation of the Portuguese Speaking Test", **Hispania**, Vol. 72, pp. 641-651, 1990.

[11] S.J. Yates, "Gender, Language and CMC for education", **Learning and Instruction**, Vol. 11, pp. 21-34, 2001.