UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

Research Notes

Issue 45 / August 2011
A quarterly publication reporting on research, test development and validation

Printed in the United Kingdom by Océ (UK) Ltd.

Research Notes

Contents

The effect of context visuals on L2 listening comprehension: Ruslan Suvorov	**2**
Issues and challenges in using English proficiency descriptor scales for assessing school-aged English language learners: Eunice Eunhee Jang, Maryam Wagner and Saskia Stille	**8**
Theoretical basis and experimental application of an auto-marking system on short answer questions: Xiangdong Gu, Fanna Meng and Wei Xiao	**14**
Teacher learning on the Delta: Simon Borg	**19**
Updates on conferences and other events	**25**

Editorial notes

Welcome to issue 45 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

This issue includes some of the papers presented at the Language Testing Research Colloquium (LTRC) in 2010, which was hosted by University of Cambridge ESOL examinations. The conference theme was *Crossing the threshold: Investigating levels, domains and frameworks in language assessment*.

We start with the paper by Ruslan Suvorov whose Master's thesis was awarded the 2009 Caroline Clapham IELTS Masters Award. Suvorov's paper, which is based on this thesis, investigates the effect of two types of context visuals (a single photograph and a video) on listening test performance. It concludes with the finding that the use of a video stimulus had a negative impact on the students' performance and draws implications for further research. Eunice Jang, Maryam Wagner and Saskia Stille discuss Steps to English Proficiency (STEP), a framework for assessing English language skills of school-aged immigrant children in Canada. They highlight the key issues and challenges in the development and validation of English proficiency descriptor scales. They also stress that the implementation of such a framework in an education context requires attention to both the content of language performance descriptors and the practices of language assessments in schools. If your interests lie in automated marking of responses to short answer questions, we draw your attention to the paper by Xiangdong Gu, Fanna Meng and Wei Xiao. The authors, who developed an auto-marking system, discuss the factors which affected the marking correctness of the system and summarise three experiments which led to a high level of marking correctness. The final paper, written by Simon Borg, investigates the impact of Delta, one of the Cambridge ESOL Teaching Awards, on teachers' learning and their teaching practices after the completion of the course. The issues discussed are relevant to language teacher educators and also contribute to the existing literature on in-service language teacher education.

We finish this issue by reporting on the events Cambridge ESOL has supported. Martin Nuttall from the ALTE Secretariat reports on ALTE events and Fiona Barker provides an update on the English Profile Project (EPP). Ardeshir Geranpayeh briefs us on the annual meeting of the National Council on Measurement in Education (NCME) which took place in New Orleans in April 2011.

The effects of context visuals on L2 listening comprehension

RUSLAN SUVOROV, IOWA STATE UNIVERSITY, AMES, USA

This paper is based on a Master's thesis supervised by Dr Volker Hegelheimer and submitted to Iowa State University (USA) in 2008. This thesis was awarded the 2009 Caroline Clapham IELTS Masters Award. A version of this paper has been published in the Technology for Second Language Learning (TSLL) Conference Proceedings (Suvorov 2009).

Introduction

Visuals have been used in second language (L2) teaching and testing for many decades. However, the role of visual support in assessing L2 learners' listening comprehension is not well understood. Research has investigated the role of visuals in L2 listening comprehension, but the little research on the effect of visual aids in L2 listening tests has been inconclusive (Buck 2001, Ockey 2007) and sound theoretical perspectives on this issue remain absent (Ginther 2001).

Researchers tend to agree that video is more authentic than audio in terms of displaying the context, discourse, paralinguistic features, and culture that should be relevant to comprehension (Coniam 2001). However, in spite of the putative merits of video, empirical research comparing audio and video modes in listening tests has been inconclusive (Buck 2001, Chung 1994, Coniam 2001). While some studies showed that visuals can improve students' performance on listening tests (e.g. Ginther 2002), others showed no facilitative effect, or in some cases even a detrimental effect, of visuals on test takers' listening comprehension (Coniam 2001, Gruba 1993, Ockey 2007). In view of the readily available authoring possibilities for listening tests today, test developers are regularly faced with the choice of including visuals in listening tests, and, therefore, further research is needed to better understand the role of visual aids in L2 listening tests.

This study investigated the role of visual support, operationalised as a single photograph and a video, in an L2 listening test for English learners. In particular, the study examined whether a difference appeared in students' performance on three parts of listening tests: one part accompanied by photographs, one with video, and one without any visual aids. This investigation began with a review of what constitutes listening comprehension.

Views of listening comprehension

Researchers have proposed numerous definitions of listening comprehension (Brett 1997, Rubin 1995). While early definitions considered linguistic sound as the means of conveying information in spoken discourse (Lado 1961), later views of listening have focused on both verbal and non-verbal stimuli (Coakley & Wolvin 1986, Rubin 1995). For example, Gruba (1997) and Ockey (2007) describe the process of listening comprehension as affected by the rate of speech, prosody, accent, phonology, and hesitations, as well as rhetorical signalling cues and the listeners' background knowledge. Some researchers argue that listening comprehension should include not only verbal input but also non-verbal components because in most real-life communicative situations, verbal information is accompanied by visual information (Buck 2001, Wagner 2007).

Visual information is believed to have a number of advantages for listening comprehension:

1. Seeing the situation and the participants increases situational and interactional authenticity, which in some cases may aid comprehension (Buck 2001, Wagner 2007).
2. Body language, facial expressions, and gestures of the speaker can provide additional information (Buck 2001, Coniam 2001, Ockey 2007).
3. With visual input, the listener can more easily identify the role of the speaker and the context (Baltova 1994, Gruba 1997, Rubin 1995).
4. Visual elements can activate the listener's background knowledge (Ockey 2007, Rubin 1995).

On the other hand, because listening involves 'making sense' (Rubin 1995:151) of the received input, the process draws upon the listener's cultural and educational background knowledge. Therefore, if the visual input does not fit into the listener's cultural expectations or background knowledge, one might hypothesise that visuals could confuse the listener and impede listening comprehension. Even though visual information seems to play an important role in oral communication, it is not clear exactly how listeners make use of various visual clues available in the process of communication. It has been suggested, however, that visuals can also be distracting or misleading when there is little or no relationship between what is said and what is shown (Rubin 1995). In other words, it is not clear that visuals can always be assumed to help comprehension.

Furthermore, the role of visual information in construct definitions of L2 listening ability in assessment is not clear and is often ignored by researchers (Buck 2001, Gruba 1993). Many test developers avoid using video in listening tests because of the possible construct irrelevant variance that they might produce in test scores (Progosh 1996). The concern is whether listening tests with visuals measure what they purport to measure, i.e. listening comprehension of language, or whether they measure some other aspects that may affect test takers' scores such as interpretation of non-linguistic meaning in visuals. Thus, many researchers claim that use of visuals in listening assessment requires careful analysis of the validity, usability, and reliability (Gruba 1997, Ockey 2007, Wagner 2007).

Buck (2001), for example, argues that in L2 listening tests aiming to measure test takers' ability to comprehend aural input rather than their ability to engage in interaction, test developers should be advised to avoid the use of video and to present a still image of the context instead. The majority of researchers, however, argue for the inclusion of non-verbal components in the construct definition of L2 listening ability, claiming that non-verbal information is an integral part of interpersonal communication in many real-life situations (e.g. Ockey 2007, Progosh 1996, Wagner 2007). They assert that the exclusion of non-verbal information from listening tests might threaten their validity (Progosh 1996, Wagner 2007), and, therefore, the listening construct in most cases needs to include the ability to obtain information from visual clues and even the ability to take notes (Ockey 2007).

The agreement on whether visuals should be included in or excluded from the construct definition of L2 listening ability can possibly be reached if we allow for the existence of different construct definitions of L2 listening ability. If the purpose of a listening test is to measure students' ability to comprehend academic lectures in the context of a university, where the students are present in an auditorium and can both hear and see a professor, then we can argue for the inclusion of visual information in the construct definition of L2 listening ability that is being measured. However, if the purpose of a listening test is to measure students' ability to understand phone conversations that are aural-only communicative situations in which interlocutors do not exchange any visual information, we can argue that the listening test and the construct definition of L2 listening ability being measured by such a test must exclude any visual information. Therefore, a decision to include or exclude visuals from the construct definition of L2 listening ability measured by a listening test should depend on the purpose of the listening test and the communicative situation presented in the test. Leaving aside the fact that many listening tests are intended to assess a construct that is relevant across more than one context, a question for test developers who wish to include visuals in listening tests is what type of visual to include.

Types of visuals

Not all visuals are the same and, accordingly, they may have different effects on listening comprehension. The two main types of visuals defined in L2 studies are *context* (or situation) visuals and *content* visuals (Bejar, Douglas, Jamieson, Nissan & Turner 2000, Ginther 2002). Context visuals provide information about the context of the verbal exchanges, such as the participants and the setting (e.g. a photo that depicts two people talking to each other in a classroom). Content visuals depict important content of the verbal interaction (e.g. a photo of Leonardo DaVinci's *Mona Lisa* accompanying a lecture on this painting). They can be classified into four groups: content visuals replicating the audio stimulus, content visuals illustrating the audio stimulus, content visuals organising information in the audio stimulus, and content visuals supplementing the audio stimulus. Bejar et al (2000) maintain that the first three types of content visuals may facilitate the comprehension of the oral stimulus, whereas the last type of content visuals may make it harder.

The effect of visuals on listening comprehension may depend on their meaning and purpose. Visuals become facilitative when the language learner can interpret their meaning correctly (Chung 1994), but they can be distracting when they decorate the text and do not convey any meaningful information (Schriver 1997). Thus, when deciding how visuals affect listening comprehension, it is important to make a distinction between context visuals and content visuals as they provide different types of information.

Visuals in L2 listening tests

In view of the content–context distinction as well as the distinction one might make between video and still images, listening tests can employ five possible modes of input: audio-only, context-only images, context-only video, content images, and content video. As the effect of different types of visuals on listening comprehension is not exactly clear (Coniam 2001), it is important to investigate the role of visuals in L2 listening tests and whether the inclusion or exclusion of different types of visuals (i.e. images or video) from the listening tests can have an impact on test takers' scores.

Several studies investigating the use of visuals in listening tests have been carried out during the last two decades (e.g. Coniam 2001, Gruba 1993, Jones 2003). However, this research has not been sufficient to provide clear results on the role of visual support in testing L2 listening. Specifically, studies are needed comparing audio-only listening tests with tests that include images and video, as well as comparative studies of different types of visuals (i.e. context and content visuals) and their effect on test takers' performance (Ginther 2002, Ockey 2007).

The few comparative studies that examined the effect of visual support on test takers' performance on L2 listening tests focused almost exclusively on context visuals. The findings of Ginther's (2002) study that employed both types of visuals suggested that the effect of visuals depended on the text types. Specifically, content visuals in mini-talks were found to be facilitative, whereas context visuals had a debilitating effect in mini-talks, no effect in dialogues/short conversations, and a facilitative effect in academic lectures. Other studies that involved only context visuals (e.g. Coniam 2001, Gruba 1993, Ockey 2007) revealed neither facilitative nor detrimental effects of visuals on L2 learners' performance on listening tests. Thus, the results of the existing research on the use of visual aids in L2 listening tests appear to be inconclusive, especially regarding context visuals that were found to be both facilitative and detrimental for test takers' performance.

In order to better understand the effect of visuals on students' performance on L2 listening tests, it is important to know if and to what extent test takers use the visual information presented to them during L2 listening tests (Wagner 2007). Additionally, individual differences among participants, including their preferences of visual aids, might have an impact on their performance on L2 listening tests. Progosh's (1996) study, for example, looked at students' preferences of visual support for listening comprehension and found that 91.9% of the learners of English as an L2 (ESL) preferred video listening tests to audio-only tests. Does it mean that students who prefer

video-mediated listening tests to audio-only listening tests will actually benefit from the visual aids and receive higher test scores? Further research is needed to address this issue.

This study addressed the following three research questions:

1. Is there a statistically significant difference among types of visual input – namely a single photograph, video, and audio-only format – in an L2 listening test in terms of their effect on L2 test takers' performance?
2. Is there a statistically significant difference between text types – namely a dialogue and a lecture – in an L2 listening test in terms of their effect on L2 test takers' performance? If there is a difference, does the effect of visuals on test takers' performance depend on text types?
3. Do test takers who prefer a particular type of visual aid perform statistically significantly better on the part of the test with their preferred type of visual than on the parts of the listening test with other types of visual input?

Methodology

The questions were addressed in a quantitative research project that used a within-subjects experimental design. The use of a within-subjects design was chosen rather than a between-subjects design to avoid error variance associated with individual differences of test takers (see the Glossary for terminology).

The quantitative data consisted of item scores for the computer-based listening test from 34 non-native learners of English enrolled in three ESL listening classes at a public Midwestern university in the USA. Additionally, the quantitative data included the participants' responses to a post-test questionnaire concerning their preferences of visual stimuli in the listening test. The independent variables measured throughout the experiment were types of visual input (video-mediated part, photo-mediated part, and audio-only part of the listening test) and text types (dialogue and lecture), as well as the participants' preferences of visuals in the listening test (video, photograph, or audio only). The dependent variables consisted of participants' scores on each of the three parts of the listening test.

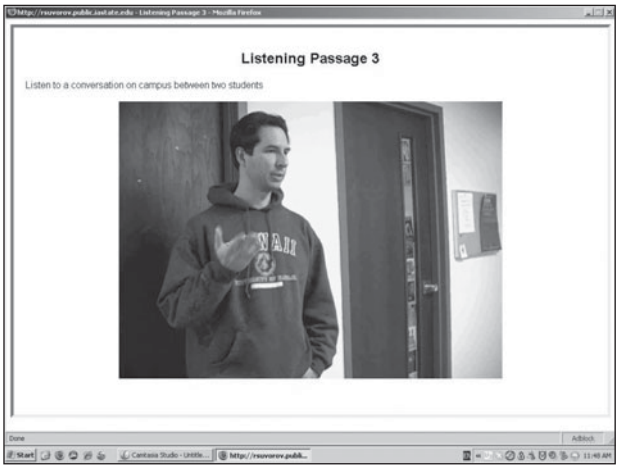
Participants

The participants were non-native speakers of English enrolled in three ESL listening classes at a large public Midwestern university: one high-level listening class with students from an Intensive English Program (IEP) group, which is a pre-university programme, and two listening classes with students enrolled in regular university classes. The overall English proficiency level of IEP students was considered to be slightly lower than that of the students in the listening class on the basis of their TOEFL scores.

A total of 34 students participated in the study, including 12 students from the IEP group, 13 students from one listening class, and nine students from the other listening group. The majority of the students were 18 to 20-year-old native speakers of Chinese. Out of 34 participants, nine were females and 25 were males. Most of the participants had been in the USA only for several months, and only five students had lived in the USA for a year or more.

Materials

The researcher designed a computer-based listening test that consisted of six listening passages and 30 multiple-choice questions (five questions for each passage) and that lasted for 44 minutes. Visual 1 provides a screenshot of the listening section with a photograph. The online version of the test is available at http://rsuvorov.public.iastate.edu/Listening_Tests/welcome.html.



Visual 1: A screenshot of one page of the listening test

Each listening passage (LP) utilised one of two text types: a dialogue between two college students or a professor and a student (D) or a short academic lecture given by a university professor (L). In addition, the researcher used one of the three types of visual input in the test: a single photograph, video, or no visuals (i.e. audio-only format). Table 1 outlines the structure of the listening test. According to the classification of visuals proposed by Bejar et al (2000) and Ginther (2002), only context visuals were used in the listening test.

Table 1: Structure of the listening test

Audio-only part (A)		Photograph part (P)		Video part (V)	
D1	L1	D2	L2	D3	L3
LP1	LP4	LP6	LP2	LP3	LP5

The texts of the listening passages, the length of which varied from 2.5 to 3.5 minutes, covered topics in Journalism, Linguistics, Biology, Sports and Nutrition, and History that do not require prior specialised knowledge in those areas. Test takers had 12 seconds in between questions to answer multiple-choice items on a computer screen. They could only hear the questions and choices and then had to choose the letter that corresponded to the best answer. Taking into account that it was a test of L2 listening comprehension, the researcher decided not to display the content of questions and answers on the screen in written form as reading skills might have come into play and affected test takers' performance and the validity of inferences about listening ability.

Five types of multiple-choice questions were used in the listening test: which is true (e.g. According to the passage, which of the following is true about pyramids?), exception (e.g. According to the passage, all of the following describe cacti EXCEPT), inference (e.g. What can be inferred about protein?), details (e.g. What is the student's main problem?), and purpose (e.g. Why does the man want to talk with the woman?).

Besides the listening test, a pre-test questionnaire and a post-test questionnaire were created. The pre-test questionnaire, consisting of 14 questions, was used to obtain information about participants' background. The post-test questionnaire that consisted of 15 open-ended questions was used to get the participants' feedback, specifically their opinions about the usefulness of visuals in the listening test. One of these questions, which was used to answer the last research question, asked test takers to indicate the type of visual input they preferred in the listening test: video, photograph, or audio-only.

Procedures

Two professors in the Department of English at the university evaluated the appropriateness of the listening test for the proficiency level of the test takers. Additionally, the researcher conducted a pilot study to check the effectiveness of procedures related to test administration, clarity of instructions and questions, quality of audio and video recording, appropriateness of listening passages, and time constraints of the study. Three international students whose overall profiles were similar to the profiles of the students in the main study participated in the pilot study. The main finding of the pilot study was that due to some technical issues related to the speed of the internet connection, the listening test should be administered locally from DVDs rather than online.

Before taking the listening test in a computer lab, students filled out a pre-test questionnaire. Because listeners, unlike readers, do not have the option of reviewing the information that has been presented to them (Thompson 1995), the participants were given paper for taking notes during the test. Note-taking allowed test takers to note down main ideas or facts from the listening passages that they could later use for answering questions. After participants finished the listening test, they were asked to fill out the post-test questionnaire. Their responses to a question about their preferences of visuals in the listening test (whether they preferred a video, photograph, or audio-only format of the listening test) were used to answer the third research question.

Data analysis

The item data from 34 participants was analysed and the reliability of the participants' scores on each section of the test was calculated. To answer the first research question, descriptive statistics for each part of the test were calculated and the ANOVA procedure with visual input types as the independent variable was used followed by the Tukey-Kramer test for post-hoc comparison. To answer the second research question, the researcher calculated descriptive statistics for two test sections defined by text type and conducted a one-sample t-test with text type as the independent variable. Then, an ANOVA procedure with a 3x2 factorial design (types of visual input by text types) was conducted followed by the Tukey-Kramer test for post-hoc comparison. To answer the third research question, the ANOVA procedure was used to determine whether the test takers' scores on the part of the test with their preferred type of visual input (which the participants self-reported in the post-test questionnaire) were statistically significantly higher (at $p<.05$) than their scores on the parts of the listening test with other types of visuals. In this analysis, the independent variable was the participants' scores on the preferred type of visuals in the listening test (i.e. video, photograph, or audio only) and the dependent variable was their scores on each of the other two parts of the listening test.

Results

Internal consistency reliability (KR-20) of the listening test scores was .70, which, considering the relatively small number of participants in this study, is acceptable. The KR-20 of scores on the audio-only part of the listening test (10 items) was .54, the KR-20 of scores on the part with a single photograph (10 items) was .63, and the KR-20 of scores on the video-mediated part (10 items) was .39. The internal consistency reliability of scores on the listening passages with dialogues was .58 (15 items) and the reliability of scores on the listening passages with lectures was .56 (15 items). The results of the ANOVA indicated that there was a statistically significant difference ($p<.05$) among mean scores for the test parts defined by types of visual input and for parts defined by text types. The ANOVA procedure with 3x2 factorial design (types of visual input by text types) showed that test takers' scores on items associated with video-mediated lecture were significantly lower than their scores on the other parts of the listening test. Overall, the students' performance on the part of the listening test that contained their preferred type of visual aid was not statistically significantly higher than their performance on the other parts of the test.

Differences related to visuals

The first research question addressed the differences in mean scores on the three parts of the test: audio, photograph, and video. To answer this question, test takers' scores on the three parts of the test were compared. Table 2 presents descriptive statistics for each part based on the type of input for 34 participants.

Table 2: Descriptive statistics for types of visual input (n=34)

Type of visual input	Number of items	Mean	SD
Audio-only	10	6.35	1.98
Photograph	10	6.32	2.25
Video	10	5.06	1.74

The ANOVA procedure followed by the Tukey-Kramer test for post-hoc comparisons found the difference between mean scores for the audio-only part of the listening test and for the part with photographs was not significant ($p=.997$).

However, significant differences were found between the means for audio-only and video-mediated parts (p=.006), as well as between the means for photo vs. video-mediated parts (p=.008).

On the basis of this data analysis, the answer to the first research question was: yes, there was a statistically significant difference in test takers' scores when listening passages with different types of visuals were used as input. Specifically, the mean score for video-mediated passages was significantly lower than mean scores for audio-only listening passages and listening passages with photographs.

Differences related to text types

The second research question addressed the differences between text types (i.e. dialogue and lecture) and their effect on test takers' performance on test items across different types of visuals. Table 3 presents the results of descriptive statistics for text types.

Table 3: Descriptive statistics for text types (n=34)

Text type	Number of items	Mean	SD
Dialogue	15	10.12	2.47
Lecture	15	7.62	2.65

A one-sample t-test was run to determine if the mean difference in students' performance on dialogues and lectures was statistically significant. The results revealed that test takers' performance on dialogues was significantly better than on lectures (p<.001).

Due to the evidence for statistically significant difference between text types, the ANOVA procedure with 3x2 factorial design (types of visual input by text types) was used to investigate a potential interaction between types of visuals and text types for the 34 participants. According to the results of the ANOVA, mean scores for the video-mediated lecture items were significantly lower than mean scores for items on the other five listening passages at the p<.05 level. However, the differences between mean scores for the video-mediated dialogue and other four listening passages with audio-only format and a photograph were not statistically significant (with p-values varying from .756 to .969) and, in fact, students performed slightly better on the video-mediated dialogue than on the audio-only lecture and the lecture with a photograph.

Thus, with regard to the second research question, the data analysis suggested that the effect of visuals on students' performance depended on the text type: while the use of video in dialogues did not have any effect on the scores, the use of video in lectures had a detrimental effect on students' performance.

Test takers' preferences vs. performance on visuals

The last research question intended to determine whether the participants of the study performed statistically significantly better on the part of the listening test with their preferred type of visual input than on the other parts of the test. The results of participants' responses to the question in the post-test questionnaire about their preferences of visuals in the listening test are given in Table 4.

Table 4: Test takers' preferences of visuals in the listening test

	Preferred type of visual input			
	Audio-only	**Photograph**	**Video**	**Total**
Number of test takers	15	7	12	34

The ANOVA procedure was used to determine whether there was a statistically significant difference (p<.05) between the test takers' scores on the part of the test with their preferred type of visual input and their scores on each of the other two parts of the listening test. The results of the analysis revealed that overall the students did not perform better on the part of the test with their preferred type of visual stimulus than on the other parts of the listening test. However, the test takers who preferred the audio-only part of the listening test scored statistically significantly higher (at p=.008) on this part than on the video part of the listening test. It is possible that the participants who preferred the audio-only format of the listening test were auditory learners who could get easily distracted by visuals and, therefore, performed significantly better on the audio-only part of the listening test.

Thus, on the basis of this data analysis, the answer to research question 3 was inconclusive: although overall the test takers who preferred a particular type of visual aid did not perform statistically significantly better on the part of the test with their preferred type of visual than on the parts of the listening test with other types of visual input, the participants with a preference for the audio-only format performed significantly better on the audio-only part of the listening test than on the video part.

Limitations

These results provide some interesting findings pertaining to the use of visuals in listening tests. However, they need to be interpreted in view of the limitations in the research. The first major limitation of this study was revealed by item analysis of the 30 multiple-choice questions used in the listening test. The results of this item analysis indicated that four out of 30 items (one item from the audio-only part, one item from the photo-mediated part, and two items from the video-mediated part of the listening test) had negative item-total correlations, which means that students who did poorly on the test overall tended to answer these four questions correctly. The deletion of these items would have raised the internal consistency reliability (KR-20) of the listening test up to .71–.72. Another way to deal with these four items would be to modify them to improve the overall internal consistency reliability of the test.

The authenticity of the listening test used in this study is also a concern. As the audio texts used in this study were designed and produced by the researcher for the study, they might lack authenticity, which according to Gruba (1997) is important in listening tests and, thus, authentic texts (i.e. the texts with a high 'degree of correspondence of the characteristics of a given language test task to the features of a T[arget] L[anguage] U[se] task' (Bachman & Palmer 1996:23)) are generally preferable.

Finally, some researchers argue that multiple-choice tests are less effective than short-answer tests or tests requiring an extended answer (Hearst 2000). Unlike short-answer tests, multiple-choice tests 'lend themselves to test-taking strategies, which do not evaluate the student's understanding of the question' (Hearst 2000:31). Therefore, as the listening test designed for this study consisted only of multiple-choice questions, the results of the study could have been different, had the listening tasks required constructed responses.

Implications for further research

The main finding of this study was that the type of visuals used in this ESL listening test affected participants' scores. The magnitude of the impact of visuals on students' performance depended on the types of visuals used. The use of a single photograph in one part of the listening test as compared to the part of the listening test without any visual support did not make any significant difference in test takers' scores; however, the use of video stimulus had a negative impact on students' performance. As this study involved only single photographs, further research is needed to evaluate the effect of multiple photographs in a listening test on test takers' performance.

In addition, different text types also appeared to have an effect on students' results on the listening test. While in dialogues the use of photographs and video did not affect participants' performance, in lectures the use of video appeared to be detrimental. The use of photographs did not seem to make any difference. Therefore, more research is needed to determine whether the use of video with different text types affects students' performance on L2 listening tests differently.

Furthermore, except for those who preferred the audio-only format, participants did not score statistically significantly higher on the part of the test with their preferred type of visual input than on the other parts of the listening test. In other words, the participants' preferences of photographs and video did not correspond with their performance on these types of visuals. If future studies corroborate this finding, further research would be necessary to determine the reasons why the students who prefer a photograph or a video do not perform better on the listening tests that use these visuals. These reasons might include but not be limited to individual differences, such as the role of cognitive load within the test takers' visual and acoustic information processing systems (Mayer 1997), learning styles of the participants, and their L2 proficiency levels.

Finally, as only context visuals were used in this study, a comparative research study on the interaction of content and context photographs and video with different text types is required to examine their effect on test takers' performance on L2 listening tests. Additionally, an interesting approach would be to use 'media inclusion' (Zenisky & Sireci 2002:348), i.e. use graphics, video, and audio within an item or set of items in an L2 listening test. Such multimedia can be employed for better illustration of a particular context, visualisation of a problem, or evaluation of a specified construct. The findings of a research study that uses visuals not only in listening passages but also in test items would greatly contribute to the understanding of the roles visuals play in listening comprehension and testing.

Glossary

ANOVA (or Analysis of Variance) – a statistical test used for comparing the means of several groups and determining whether there are any significant differences among them.

Between-subjects design – a type of experimental design in which different groups of subjects are exposed to the same treatment.

Tukey-Kramer test – a multiple comparison procedure that usually accompanies the ANOVA test to determine which means differ significantly from one another.

Within-subjects design – a type of experimental design in which the same group of subjects is exposed to different treatments.

References and further reading

Bachman, L and Palmer, A (1996) *Language testing in practice: Designing and developing useful language tests*, Oxford: Oxford University Press.

Baltova, I (1994) The impact of video on the comprehension skills of core French students, *Canadian Modern Language Review* 50 (3), 507–531.

Bejar, I, Douglas, D, Jamieson, J, Nissan, S and Turner, J (2000) *TOEFL 2000 listening framework: A working paper*, Princeton: Educational Testing Service.

Brett, P (1997) A comparative study of the effects of the use of multimedia on listening comprehension, *System* 25 (1), 39–53.

Buck, G (2001) *Assessing listening*, Cambridge: Cambridge University Press.

Chung, U (1994) *The effect of audio, a single picture, multiple pictures, or video on second-language listening comprehension*, unpublished PhD dissertation, University of Illinois at Urbana-Champaign.

Coakley, C and Wolvin, A (1986) Listening in the native language, in Wing, B (Ed.) *Listening, reading, and writing: Analysis and application*, Middlebury: Northeast Conference, 11–42.

Coniam, D (2001) The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study, *System* 29, 1–14.

Ginther, A (2001) *Effects of the presence and absence of visuals on performance on TOEFL CBT listening-comprehension stimuli*, Report 66, Princeton: Educational Testing Service.

Ginther, A (2002) Context and content visuals and performance on listening comprehension stimuli, *Language Testing* 19 (2), 133–167.

Gruba, P (1993) A comparison study of audio and video in language testing, *JALT Journal* 15 (1), 85–88.

Gruba, P (1997) The role of video media in listening assessment, *System* 25 (3), 335–345.

Hearst, M (2000) The debate on automated essay grading, *IEEE Intelligent Systems* 15 (5), 22–37.

Jones, L (2003) Supporting listening comprehension and vocabulary acquisition with multimedia annotations: The students' voice, *CALICO Journal* 21 (1), 41–65.

Lado, R (1961) *Language testing: The construction and use of foreign language tests*, London: Longman.

Mayer, R (1997) Multimedia learning: Are we asking the right questions?, *Educational Psychologist* 32 (1), 1–19.

Ockey, G (2007) Construct implications of including still image or video in computer-based listening tests, *Language Testing* 24 (4), 517–537.

Progosh, D (1996) Using video for listening assessment: Opinions of test-takers, *TESL Canada Journal* 14 (1), 34–44.

Rubin, J (1995) The contribution of video to the development of competence in listening, in Mendelsohn, D and Rubin J (Eds) *A guide for the teaching of second language listening*, San Diego: Dominie Press, 151–165.

Schriver, K (1997) *Dynamics in document design: Creating text for readers*, New York: John Wiley & Sons.

Suvorov, R (2009) Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format, in Chapelle, C A, Jun, H G and Katz, I (Eds) *Developing and evaluating language learning materials,* Ames, IA: Iowa State University, 53–68.

Thompson, I (1995) Assessment of second/foreign language listening comprehension, in Mendelsohn, D and Rubin J (Eds) *A guide for the teaching of second language listening*, San Diego: Dominie Press, 31–58.

Wagner, E (2007) Are they watching? Test-taker viewing behavior during an L2 video listening test, *Language Learning and Technology* 11 (1), 67–86.

Zenisky, A and Sireci, S (2002) Technological innovations in large-scale assessment, *Applied Measurement in Education* 15 (4), 337–362.

Issues and challenges in using English proficiency descriptor scales for assessing school-aged English language learners

EUNICE EUNHEE JANG, MARYAM WAGNER AND SASKIA STILLE, ONTARIO INSTITUTE FOR STUDIES IN EDUCATION, UNIVERSITY OF TORONTO

Introduction

Increasing globalisation and mobility of people across countries and borders has resulted in shifting demographics and interaction between people, cultures and languages across the world. This change is reflected in Canada which has one of the highest immigration rates per capita in the world. In the province of Ontario, Canada, over 20% of the student population comprises English language learners (ELLs). These students have to catch up with their same-age peers not only in acquiring curriculum content, but also in developing academic language proficiency which involves increasingly more complex language skills at each grade level. These challenges underlie some of the achievement and opportunity gaps that arise between ELLs and their English-speaking peers in K-12 schools in Ontario (see Jang, Kim, Gu, Zhang, Wu & Wagner 2009). Further, these circumstances demand more systematic approaches to ensuring that all students receive the language support necessary to gain English language proficiency and achieve academic success in school.

Steps to English Proficiency (STEP) is a language assessment framework for use in K-12 schools. It was developed by English as a second language (ESL) content experts and teachers to support a policy initiative by the Ontario Ministry of Education to improve the assessment, tracking, and support of ELLs by classroom teachers. In this paper, we introduce the STEP language assessment framework, discuss key issues related to the development of English proficiency descriptor scales for school-aged ELLs, and reflect on challenges to validating the scales. In particular, we address how the implementation of a language assessment framework for use in educational contexts requires attention to not only the content of performance descriptors, but also to the practices and processes of language assessment in schools, and the challenges of aligning language proficiency performance scales with curriculum and instruction in schooling. We begin our discussion by discussing STEP in the broader context of proficiency descriptors-based assessment systems.

Development of STEP language proficiency descriptor scales

The development of performance-based descriptors supports an increasingly utilised, policy-supported practice of alternative assessment to assist ELLs in K-12 schools. The wide use of alternative assessments in classrooms reflects a shift and expansion in the scope and purpose of assessment, from testing of discrete knowledge and skills to classroom authentic assessment of language proficiency in a specific language-in-use context where many students are learning English as an additional language (Brindley 2001, Chalhoub-Deville 2003, Shepard 2002). Authentic classroom assessments offer opportunities to observe and evaluate students through tasks that are embedded in the curriculum (Darling-Hammond and Snyder 2000). Thus, the shift represents an integration of assessment with teaching and learning; a process which may be further facilitated through the use of language assessment proficiency scales such as the Canadian Language Benchmarks (CLB), American Council on the Teaching of Foreign Languages (ACTFL), World-Class Instructional Design and Assessment (WIDA) English Language Proficiency Standards, and the Common European Framework of Reference for Languages (CEFR). The Ministry of Ontario in Canada recently contributed to

this list of exemplars through the development of the STEP assessment framework responding to a call to provide teachers with tools to consistently assess and track school-aged ELLs' English proficiency on a periodic basis. In developing the STEP descriptors-based framework, educators prioritised the following principles of assessment, articulating its need to:

- be fully integrated into teaching and learning
- be aligned with grade-appropriate curricular expectations
- serve formative purposes of assessment by facilitating the development of school-aged ELLs' English proficiency
- be based on authentic learning tasks and teachers' observations.

The STEP framework consists of three sets of English proficiency descriptor scales, each comprising six levels or steps, for Reading and Responding, Writing, and Oral Communication skills for each of four grade clusters (Grades 1–3, 4–6, 7–8, and 9–12). The operational construct of each skill is guided by curricular expectations and theoretical underpinnings. For example, the Writing skill emphasises writing as process (e.g. Seow 2002, Tribble 1996) rather than product (e.g. Pincas 1982), which at the secondary level involves the following elements: (1) engaging in prewriting to generate and organise ideas and information; (2) incorporating a variety of text forms and features in writing; (3) writing with fluency, using a variety of sentence structures and transition words, and (4) revising for content and clarity.

The Oral Communication skill is defined in terms of communicative competence of linguistic, sociolinguistic, and strategic components (Bachman 1990, Canale & Swain 1980). Note that Oral Communication includes both listening and speaking abilities, which reflects the fact that listening ability is integrated into the Oral Communication skill in the K-12 curriculum. Listening is further operationalised into four main elements. For example, in the secondary panel (grades 9–12), these skills are delineated as: listening for a variety of purposes; using comprehension strategies; analysing and interpreting oral texts; and, making connections. At the secondary level, the speaking ability defined within the Oral Communication continua includes: (1) using language specific to the purpose, (2) using verbal skills and non-verbal cues; and (3) using language with clarity and coherence.

The Reading and Responding skill is operationalised according to the interactive model that emphasises the importance of both the text and the reader in the process of making meanings of incoming textual information, based on the information that the reader brings to the text (Grabe 1991, Stanovich 1990). Specifically, the Reading and Responding skill includes three elements, which at the secondary level are defined as: (1) understanding and responding to texts using strategies (Meaning), (2) using text features, text forms, and style to construct meaning (Form and Style), and (3) reading and understanding familiar and unfamiliar words and phrases, and expanding vocabulary (Fluency).

Each continuum consists of three components. 'Student Profile' provides a diagnostic summary of students' language proficiency at the end of each Step. 'Element' describes language concepts and skills that make a connection to curriculum. The descriptors of 'Observable Language Behaviours' capture distinct language behaviours that teachers can observe and evaluate in the context of their curricular teaching and learning activities. Curricular-specific examples of evidence are provided to illustrate learning behaviours specific to oral, reading and writing modalities. Table 1 illustrates a sample continuum chart which highlights how, together, these components provide teachers with the necessary information to evaluate students' current English language proficiency, and also guide their future instruction and planning to help students' language learning progression.

Each descriptor aims for independence (Alderson 1991, North & Schneider 1998, Turner & Upshur 2002) to capture distinct levels of student performance. The descriptors are listed on tracking sheets that teachers use to record individual learners' development, and to identify future learning goals or objectives to support learners' continued progress. English literacy development (ELD) students whose first languages are other than English or are a variety of English significantly different from that used for instruction in Ontario schools, are accommodated with separate sets of descriptors for Reading and Responding, and Writing for

Table 1: Sample secondary writing continuum

Writing Summary Profile – STEP 3						
Students communicate ideas and information, using English words, phrases, sentences, and/or L1. They bring age-appropriate concepts of literacy in their first language to writing in English. Students write in a variety of text forms for different purposes and audiences. They write a variety of simple and compound sentences, using familiar and pre-taught subject-specific vocabulary. They revise to develop information and ideas, and clarify meaning.						
	Observable Language Behaviours (OLB)					
Element	**STEP 1**	**STEP 2**	**STEP 3**	**STEP 4**	**STEP 5**	**STEP 6**
Revising Revise for content and clarity	Use teacher feedback and classroom resources to make corrections to individual words (e.g. word wall, picture dictionary, self-created dictionary)	Correct teacher- or peer-identified spelling errors by using various resources (e.g. high-frequency word lists, word family lists, picture and/or dual language dictionaries	Correct self-identified spelling errors by using various resources (e.g. high-frequency word lists, word family lists, and/or junior dictionaries	Revise to address specific writing conventions, using an editing checklist	Revise after re-reading to ensure a logical and fluent presentation of information or ideas	Cross-check a draft against writing plan to identify parts that need to be added, moved, or deleted

Steps 1–4 to reflect these students' gaps in their opportunity for literacy development. In addition to these classroom-based materials, STEP also provides initial assessment materials for assessment of newcomer students' English language proficiency.

Validation of STEP

STEP has undergone various stages of field research and revision in a process to meet these objectives during its use in classrooms. We collaborated with the Ministry of Education to gather validity evidence to support the use of STEP in Ontario classrooms. In Phase 1 (2007–08), we prioritised five validity concerns in discussion with the stakeholders in the project, focusing on the impact of STEP, construct representation and interpretability of descriptors, fairness of STEP in addressing the linguistic and cultural diversity of students, and consistency of the use of STEP. We examined those issues by seeking evidence from interviews and focus groups with 35 teachers from 25 different schools who had the opportunity to assess their students repeatedly over time after receiving training. Some of the key findings emerging from this phase of the investigation illustrated that STEP served a pedagogical and educative purpose for teachers advancing their knowledge of ELLs' English language development, and providing teachers with a common language and framework of reference (Butler & Stevens 1998). At the same time, teachers identified challenges related to increased work-load (Brindley 2001).

During Phase 1, teachers' interviews revealed that they found STEP descriptors to be relevant to the modality and interpretable in a broad sense, however they identified inconsistencies across the continua and the difficulty in distinguishing between certain STEPs. Furthermore, teachers reported that some descriptors were normalised against native speakers highlighting both construct and fairness issues related to the descriptors. Other issues related to the fairness included teachers' alerting us to the (unfair) inclusion of pronunciation in the descriptors, as well as potential misuses of STEP to identify cognitive exceptionalities.

The results from Phase 1 informed the next phase of the study. Phase 2 (2009) involved 17 item writers and ESL specialists in a summer item-writing workshop in order to improve the descriptors' linguistic and conceptual clarity and refine the descriptors according to their relevance to grade levels and target skills, based on similar work done in the development and validation of other descriptors-based language proficiency scales (e.g. North 1993, Turner & Upshur 2002, WIDA Research Brief 2007). The participants identified problems with some of the descriptors including: overloaded descriptors, lack of specificity describing colloquial and idiomatic language, descriptors not appropriate to grade and/or STEP level. Additionally, they highlighted the problematic use of unobservable cognitive processes defined in the descriptors, e.g. compensating, learning, and knowing. Perhaps one of the most important findings emerging from Phase 2 of the validation study was that although there is a need for descriptors to distinguish between language acquisition and literacy tied to the curriculum, they still need be aligned with the curriculum so that they are relevant to the classroom context.

Phase 3 (2010) field-tested revised STEP descriptors with a much larger sample of 170 teachers across Ontario. The purposes of examining descriptor quality were to flag potentially problematic descriptors and identify possible sources of problems underlying these descriptors. High agreement rates were achieved amongst the participants of the study who evaluated the descriptors for clarity and appropriateness to grade cluster, language skill, and STEP level. With respect to clarity, the participants identified three main aspects which needed to be addressed in some descriptors: inclusion of curriculum content-specific language without elaboration of related linguistic behaviours, use of ambiguous language, and inclusion of more than one competency within a skill. Participants identifying descriptors as inappropriate to a language skill were often referring to: linguistic competencies that related to more than a single skill or those that are not observable in a classroom. They also identified concerns with the inclusion of pragmatic competencies such as use of idiomatic language and humour. Descriptors not appropriate to a STEP level may have been better suited to higher/lower STEP levels or simply not observable at the specified level. This data provided fine-grained, detailed information for item writers participating in a subsequent round of descriptor revisions.

The current phase is focused on examining the stability of the STEP proficiency descriptor scales by collecting large-scale student data through collaboration with teachers across the province. The student assessment data is being used to establish empirical evidence that confirms/disconfirms the stability of the current six-step scales. Additionally, we are interested in examining the extent to which the interpretation and use of the current descriptors are fair for all students regardless of their linguistic and cultural background.

While collaborating with multiple stakeholders, including mainstream and ESL teachers, school board educators, and policy makers, we have identified several conceptual issues in the use of proficiency descriptor scales for assessing school-aged ELLs' English proficiency in school contexts. In the following section, we focus our discussion on four issues: role of proficiency descriptor scales in classroom assessment; role of stakeholders in developing and validating the scales; role of context specificity; and integration of language acquisition and literacy.

Role of proficiency descriptor scales in classroom assessment

Language proficiency scales, like STEP, have been developed in response to the need for integrating teaching, learning and assessment, guided by curricular frameworks or expectations. A range of studies, across different national contexts, report on the development of English proficiency descriptors for assessing ELLs' language development in their learning contexts (Brindley 1998, McKay 2000, Scott 2009, Scott & Erduran 2004). Teachers may use the scales to evaluate their students' language development and proficiency according to their performance on tasks

(Wigglesworth 2008, Norris, Brown, Hudson & Yoshioka 1998). When integrated into educational contexts, these tasks refer to the curricular activities that all students are doing every day during teaching and learning activities in the classroom. One of the main advantages of performance-based assessments is that they provide the opportunity for language use to be observed, and consequently rated, in a direct, authentic context (Brindley 2001, Cumming 2008, McNamara 1996, Messick 1994).

Evidence from the aforementioned multiple-phase STEP field research supports the positive roles of performance-based classroom assessments based on proficiency descriptor scales, including its potential to: advance teachers' understanding of language development across curricula; affirm teachers' role as a main agent of assessment; provide teachers with a common language and framework of reference; draw teachers' attention to the wide range of learning activities elicited by the descriptors; and inform instructional planning.

Needless to say, the key strength of using proficiency descriptor scales in classroom assessment is its capacity to be aligned and integrated into instructional activities as well as its formative and diagnostic potential to inform students' strengths and weaknesses (Norris, Brown, Hudson & Yoshioka 1998). It is this formative quality, or potential for positive washback, of proficiency descriptors that distinguishes them from traditional single-score language assessments. In fact, a further impetus for the use of proficiency descriptors is prompted by the need to inform teachers' selection of instructional strategies appropriate for supporting ELLs during subject-area instruction. With this diagnostic and formative purpose, test users, such as teachers, are active participants in language assessment, wherein they are assisted in evaluating and supporting learners' progress in developing the language skills that are critical for success in school-based learning.

Roles of stakeholders in developing proficiency descriptor scales

The development and validation of language proficiency descriptor scales typically involves various stakeholders who bring different professional knowledge and experience to the task, such as teachers, teacher educators, school administrators, content-area experts, assessment specialists, policy makers, and/or students and parents. The development and validation of STEP was based on a triangular relationship among the government agency (Ministry of Education), ESL teachers (content experts and research participants), and the researchers (or evaluators), each of whose roles was distinct and equally pivotal. The Ministry led the development of the STEP continua. Their work was motivated to serve public accountability purposes, while contracted ESL teachers were charged with the development of the STEP continua. In the subsequent validation field research, numerous ESL teachers provided input based on their experience with STEP. We, as researchers, were contracted to evaluate the validity of the STEP framework. We were and continue to be accountable to our funding body, the Ministry, but we were also guided by professional standards and ethics which inform decision making process.

With distinct roles played by different stakeholders, the researchers are often perceived as outside experts who would provide objective evidence based on scientific research. Negotiating conflicts arising from different interests and power distribution among these groups becomes inevitable. For example, one of the most challenging dilemmas we encountered is the question about the extent to which the continua are or should be linked to the curriculum, students' cognitive developmental stages, and instructional tasks. This question was debated among stakeholders. ESL teachers expressed their concern that too strong a focus on curriculum expectations would result in assessment of literacy, not overall language development. The Ministry's primary concern was to ensure that the STEP descriptors would be aligned with the school curriculum and accessible to both ESL teachers and mainstream classroom teachers, whose limited knowledge experience of ESL education is limited. The stakeholders brought their agendas and views to the debates with further questions, such as: 'Is STEP meant to be a language assessment framework or a model of language instruction through curricula?' (Fulcher 2008); 'Does STEP's link to the curriculum weaken the clarity in the construct definition of language proficiency?'; and 'How can language be assessed independent of subject-specific content? (Byrnes 2008).

Inclusion and deliberation emerged as two central elements of this democratic evaluation process to address the complexity of power relations among stakeholders, alongside the introduction of a new language assessment framework, calls for a shift towards democratic validation/evaluation (House & Howe 2000, Howe & Ashcraft 2005, Jang, Wagner & Stille 2010, Ryan 2005). Moss (1994) notes the importance of gaining consensus among stakeholders in claims of validity, suggesting a subjective quality that underlies the reliability and appropriateness of language proficiency scales. A need for reconceptualising the notion of validation alludes to the complex relationships and shared responsibility involved in ensuring that language assessment frameworks make an impact on curricula and pedagogy in K-12 schools (Bonnet 2007).

Role of context specificity in STEP

Because STEP is specifically developed to assess the learning trajectories of ELL students (both immigrants to Canada and students born in Canada whose L1 is other than English or French), its focus and scope are specific to K-12 educational content and are guided by a set of curricular expectations appropriate to different cognitive development stages, as evidenced in different descriptors per step and across grade clusters. The STEP continua describe proficiency levels in sufficient detail that teachers can make a direct link to their observation of language use. Table 2 illustrates some primary level (grades 1–3) Reading descriptors used to assess ELLs' ability to read and understand familiar and unfamiliar words and phrases.

Table 2: Some examples of primary Reading descriptors

STEP 1	STEP 2	STEP 3	STEP 4	STEP 5	STEP 6
Recognise and comprehend high-frequency words in a few contexts (e.g. calendar, class chart)	Decode unfamiliar words in highly illustrated texts by making sound-symbol connections	Determine the meaning of unfamiliar vocabulary, using root words, prefixes and suffixes (e.g. happy/unhappy)	Locate and use subject-specific key vocabulary (e.g. to complete graphic organisers)	Incorporate low-frequency vocabulary from reading into written work and oral responses	Understand most vocabulary in a variety of grade-appropriate texts

Aligning the development of English proficiency with curricular expectations involves articulating levels of language progression based on actual learner performance in schools (Byrnes 2007, North 2007). The learning environment for K-12 ELLs is not limited to language learning, but includes curricular learning. As such, English proficiency scales for school-aged children should be embedded into their specific learning goals and contexts. The scales should provide teachers the opportunity to observe linguistic performance based on their use of language during learning tasks. This context specificity requiring tight curricular alignments is different from context-neutral proficiency scales, such as the Common European Framework of Reference for Languages (CEFR). In contrast, the CEFR is a reference framework that requires contextualisation to serve as the basis for implementation (Little, personal communication). As Jones & Saville (2009:55–66) add:

> … people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring a context to the CEFR … The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context that determines the final meaning of the claim. By engaging with the process in this way, we put the CEFR in its correct place as a point of reference and also contribute to its future evolution.

By prioritising the K-12 learning context, STEP aligns descriptors with Ontario curricula and instructional activities, and embeds assessment into teachers' daily practice in order to allow teachers to directly observe and evaluate learners' language based on their performance while engaged in meaningful, classroom authentic tasks. Describing learners' linguistic abilities within the context of curriculum learning has increased the potential for STEP to have an impact on curriculum and pedagogy and support for ELLs in schools (Little 2010). For teachers, the continua not only describe learners' current level of language development, but also suggest the kinds of educational scaffolding and instruction that would support the further development of these abilities (Little 2010, Snow & Uccelli 2009).

Language development of school-aged ELLs

Another crucial issue in the development of proficiency descriptors is how language development is defined in K-12 school contexts in which it is primarily characterised in terms of oral language proficiency and literacy development. A lack of comprehensive theories about and empirical research on the relationship between oral and literacy development of school-aged ELLs present confounding challenges. In general, first-language learners start their schooling after having commenced reading and writing with well-developed oral proficiency in their (first) language. In other words, their literacy development builds on oral language proficiency (McKay 2006). In contrast ELLs show different developmental patterns, as McKay (2006:13) states:

> Foreign language learners bring a background of literacy development in their first language to their language learning. Their skills in literacy in the foreign language build on their developing first language literacy understanding and skills but are dominated by a lack of oral knowledge of the foreign language. Many young second language learners have not had an opportunity to develop first language literacy skills and are therefore learning literacy in their second language, compounding the challenge of second language learning. Hence, not only do these children not have literacy understanding and skills, they do not have oral knowledge of the new language.

The quote signifies the interrelationship between oral and literacy development. It further suggests that the developers and users of proficiency descriptor scales (to be used in K-12 school contexts) need a clear knowledge base about the trajectories of ELLs' literacy development and its interface with oral language proficiency development in order to address the holistic language learning and assessment perspectives.

As we noted, the alignment between STEP descriptors and curriculum delineates the components of literacy development in detail. Yet, the curriculum is not sufficiently detailed to the point of defining what language ability means (McKay 2006). Accordingly, teachers felt that the STEP descriptors were too tightly aligned with the curriculum, and expressed their concern that the STEP scales may be limited to assessing literacy development rather than English language proficiency. Prompted by the teachers' concern, we are currently examining the viability of various theoretical frameworks of school-aged ELLs' language development. Similar issues have been raised in discussions of the CEFR whose descriptors were developed by expert teachers (North 1993, North & Schneider 1998). If the construct of 'English language development' remains ill-defined then efforts to develop valid measures or procedures to map the trajectories of ELL students' linguistic and academic development will yield inconsistent and inconclusive data. The material consequences of this failure to define the underlying theoretical construct in a coherent and measurable way include difficulties in identifying stages in students' English language development.

Conclusion

In general, the development and use of English language assessment frameworks in classroom contexts presents

unique issues and challenges relating to the operationalisation of language proficiency growth and the assessment of these processes by teachers. Despite these challenges, language assessment frameworks have the capacity to bring curricula, pedagogy, and assessment into much closer interdependence in K-12 school contexts (Little 2010).

The implementation of the STEP framework focused attention on the unique issues of assessment in the educational context; specifically, the importance of aligning descriptors with curriculum, and the relationship between language and literacy development in school-based learning. The limitations of this short discussion are clear: these issues need to be explored through the analysis of empirical data. The next phase of research relating to STEP involves the collection of student assessment data which will provide the opportunity to explore the issues articulated in this discussion. As STEP is fully implemented and more extensive data become available, it will be possible to examine the factorial structure and theoretical coherence of the descriptors empirically. In the European context, it is likely that CEFR will be employed for similar purposes and it will no doubt be important to assess the adequacy of each instrument in mapping immigrant students' language and literacy trajectories and also the utility of each instrument in informing policy and instructional decisions for this population of students.

References

Alderson, J C (1991) Language Testing in the 1990s: How far have we come? How much further have we to go?, in Anivan, S (Ed.) *Current Developments in Language Testing. Singapore*, RELC Press, 1–26.

Bachman, L (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bonnet, G (2007) The CEFR and Educational Policies in Europe, *Modern Language Journal* 91 (4), 669–672.

Brindley, G (1998) Describing language development? Rating scales and SLA, in Bachman, L F and Cohen, A D (Eds) *Interfaces Between Second Language Acquisition and Language Testing Research*, Cambridge: Cambridge University Press, 112–140.

Brindley, G (2001) Outcomes-Based Assessment in Practice: Some Examples and Emerging Insights, *Language Testing* 18 (4), 393–407.

Butler, F A and Stevens, R (1998) *Initial steps in the validation of second language proficiency descriptors for Public High Schools, Colleges, and Universities in California: Writing*, CSE Technical Report 497, Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), 3–92.

Byrnes, H (2007) Perspectives, *The Modern Language Journal* 91 (4), 641–645.

Byrnes, H (2008) Assessing content and language, in Shohamy, E and Hornberger, N H (Vol. Eds) *Language Testing and Assessment*, New York: Springer, 37–52.

Canale, M and Swain, M (1980) Theoretical Bases of Communicative Approaches to Second-Language Teaching and Testing, *Applied Linguistics* 1 (1), 1–47.

Chalhoub-Deville, M (2003) Second Language Interaction: Current Perspectives and Future Trends, *Language Testing* 20 (4), 369–383.

Cumming, A (2008) Assessing oral and literate abilities, in Shohamy, E and Hornberger, N H (Eds) *Encyclopedia of Language and Education*, 2nd edition, Volume 7: *Language Testing and Assessment*, New York: Springer, 3–18.

Darling-Hammond, L and Snyder, J (2000) Authentic assessment of teaching in context, *Teaching and Teacher Education* 16 (5), 523–545.

Fulcher, G (2008) Criteria for evaluating language quality, in Shohamy, E and Hornberger, N H (Eds) *Encyclopedia of Language and Education*, 2nd edition, Volume 7: *Language Testing and Assessment*, New York: Springer, 157–176.

Grabe, W (1991) Current Development in Second Language Reading Research, *TESOL Quarterly* 25 (3), 375–406.

House, E R and Howe, K R (2000) Deliberative democratic evaluation in practice, in Stufflebean, D L, Madaus, G F, and Kellaghan, T (Eds) *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, Higham, MA: Lluwer Academic Publishers, 409–440.

Howe, K R and Ashcraft C (2005) Deliberative Democratic Evaluation: Successes and Limitations of an Evaluation of School Choice, *Teachers College Record* 107 (10), 2,275–2,298.

Jang, E E, Kim, Y-H, Gu, Z, Zhang, J, Wu, Y, and Wagner, M (2009) *Tracking developmental patterns of reading skill profiles among elementary school students: Application of cognitive diagnostic modelling*, symposium presented at the Canadian Society for the Study of Education (CSSE), Ottawa, 23–29 May, 2009.

Jang, E E, Wagner, M and Stille, S (2010) A Democratic Evaluation Approach to Validating a New English Language Learner Assessment System: The Case of Steps to English Proficiency, *English Language Assessment* 4, 35–50.

Jones, N and Saville, N (2009) European Language Policy: Assessment, Learning, and the CEFR, *Annual Review of Applied Linguistics* 29, 90–100.

Little, D (2010) *The Common European Framework of References for Languages: a research agenda*, paper presented at OISE/University of Toronto, Toronto, Canada.

McKay, P (2000) On ESL Standards for School-Age Learners, *Language Testing* 17, 185–214.

McKay, P (2006) *Assessing Young Language Learners*, Cambridge: Cambridge University Press.

McNamara, T F (1996) *Measuring Second Language Performance*, London: Longman.

Messick, S (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessment, *Educational Researcher* 23 (2), 13–23.

Moss, P (1994) Can There be Validity Without Reliability?, *Educational Researcher* 23 (2), 5–12.

Norris, J M, Brown, T D, Hudson, T, and Yoshioka, J K (1998) *Designing Second Language Performance Assessments*, Honolulu: University of Hawaii Press.

North, B (1993) *The Development of Descriptors on Scales of Language Proficiency*, Washington, DC: National Foreign Language Center.

North, B (2007) The CEFR Illustrative Descriptor Scales, *The Modern Language Journal* 91 (4), 656–659.

North, B and Schneider, G (1998) Scaling Descriptors for Language Proficiency Scales, *Language Testing* 15 (2), 217–263.

Pincas, A (1982) *Teaching English Writing*, London: Macmillan.

Ryan, K (2005) Democratic Evaluation Approaches for Equity and Inclusion, *The Evaluation Exchange* 11 (3), 2–3.

Scott, C (2009) Issues in the Development of a Descriptor Framework for Classroom-Based Teacher Assessment of English as an Additional Language, *TESOL Quarterly* 43 (3) 530–535.

Scott, C and Erduran, S (2004) Learning from International Frameworks for Assessment: EAL Descriptors in Australia and the USA, *Language Testing* 21, 409–431.

Seow, A (2002) The writing process and process writing, in Richards, J C and Renandya, W A (Eds) *Methodology in Language Teaching: An Anthology of Current Practice*, Cambridge: Cambridge University Press, 315–320.

Shepard, L (2002) Standardized tests and high-stakes assessment, in Gurthrie, J W (Ed.) *Encyclopedia of Education*, New York: Macmillan Reference, 2,533–2,537.

Snow, C and Uccelli, P (2009) The challenge of academic language, in Olson, D R and Torrance, N (Eds) *The Cambridge Handbook of Literacy*, Cambridge: Cambridge University Press, 112–133.

Stanovich, K (1990) Concepts of Developmental Theories of Reading Skill: Cognitive Resources, Automaticity, and Modularity, *Developmental Review* 10, 72–100.

Tribble, C (1996) *Writing*, Oxford: Oxford University Press.

Turner, C E and Upshur, J A (2002) Rating Scales Derived From Student Samples: Effects of the Scale Maker and the Student Sample on Scale Content and Student Scores, *TESOL Quarterly* 36, 49–70.

Wigglesworth, G (2008) Task and performance based assessment, in Shohamy, E and Hornberger, N H (Vol. Eds) *Language Testing and Assessment*, New York: Springer, 111–122.

World-Class Instructional Design and Assessment (WIDA) (2007) *English Language Proficiency Standards and Assessment Tools*, retrieved from www.wida.us/standards/elp.aspx

Theoretical basis and experimental application of an auto-marking system on short answer questions

XIANGDONG GU RESEARCH CENTER OF LANGUAGE, COGNITION AND LANGUAGE APPLICATION, CHONGQING UNIVERSITY, CHINA
FANNA MENG COLLEGE OF FOREIGN LANGUAGES, HENAN NORMAL UNIVERSITY, CHINA
WEI XIAO COLLEGE OF FOREIGN LANGUAGES, CHONGQING UNIVERSITY, CHINA

Introduction[1]

'Short Answer Questions are "constructed-response", or open-ended questions that require students to create an answer. Short answer items typically require responses of one word to a few sentences' (Newble & Cannon 1989:107). In this study, short answer questions (SAQs) refer to the test format of a question or an incomplete statement that expects test takers to respond in one word or a few words. Being a subjective task type, SAQs have potential for positive washback on the teaching and learning of English. However, their scoring process needs plenty of human resources and, given the subjective nature of the task type, scoring reliability cannot be guaranteed. Possibly on account of this, SAQs are probably not as widely adopted in large-scale English tests as multiple-choice questions.

In recent years, along with the rapid development of computer technology, there have been an increasing number of studies concerning the auto-marking of subjective tests. In western countries, many new assessment programmes have emerged. The most prominent are three essay-rating systems: Project Essay Grade (PEG), Intelligent Essay Assessor (IEA) and Electronic Essay Rater (E-rater). While PEG focuses only on language form and IEA only on content, E-rater is superior to these in that it focuses on both form and content (Liang & Wen 2007). According to Valenti, Neri & Cucchiarelli (2003), since E-rater was launched in 1999, it has scored more than 750,000 essays, reaching over 97% agreement with the scores awarded by human raters.

As for the study of marking SAQs, Burstein, Kaplan, Wolff & Lu (1996) carried out research on the auto-marking of short answers that consist of a few words or a short sentence. The accuracy of their marking system – Automark – reached 93–96%. However, Automark gives grades rather than specific scores to students' responses. Besides, too much time tends to be needed for pretreatment, which is a focus of subsequent efforts for improvement (Valenti et al 2003).

In China, research into the auto-marking of subjectively marked tests is still in its infancy. Gao & Yuan (2004) and Meng, Bu, Li & Gan (2005) found that two factors – key words and similar degree –affect the scoring on subjective tests. They also designed an automated assessment algorithm to imitate a human rater's scoring process, although they do not provide a detailed methodology for arriving at a score by using the two aforementioned factors. Besides, they provide an algorithm which, they claim, is suitable for any subjective task/test type (e.g. SAQs, open-ended questions or essay writing). However, if the algorithm was put into use for a specific type of a subjective test, such as SAQs, it would need to be revised according to the features of the specific test. A study reporting on development of an auto-marking model for English essay-rating, Liang (2005), mainly focuses on the statistical analysis of some superficial features such as noun phrases. Owing to the limitation of the sample (733 essays in total, 200 for model training and 533 for auto-marking), Liang's marking model is not yet applicable.

On the basis of the previous research, an auto-marking system for SAQs was designed as part of this study, employing the Theory of Single Similar Degree in Fuzzy Mathematics. The system was subsequently improved through three experiments, which are also discussed in the present paper.

Research questions

The following Research questions will be addressed in this study:

[1] This paper is part of the research project A Longitudinal Study of the CET Washback supported by the National Philosophy and Social Science Foundation of China (07BYY030) and by the National Research Centre for Foreign Language Education (MOE Key Research Institute of Humanities and Social Sciences at Universities), Beijing Foreign Studies University.

1. How can an auto-marking system be designed by using Single Similar Degree in order to mark SAQs efficiently and accurately?
2. How high is the scoring reliability of the auto-marking system? If it is not as high as expected, how can it be improved?

Theoretical basis

This section begins with a discussion of certain concepts from Fuzzy Mathematics which were crucial for the development of the auto-marking system presented in this study. It ends with an example of how a student's response to a short answer question is auto-marked.

In 1965, the publication of the paper *Fuzzy Sets* by the mathematician Zadeh (1965) signalled the birth of the discipline Fuzzy Mathematics. As a new branch of mathematics, Fuzzy Mathematics enables disciplines normally not associated with mathematics (e.g. Biology, Psychology, Linguistics) to describe and analyse the data in novel ways through the use of computer technology. For example, thanks to Fuzzy Mathematics, computers are not only able to make correct/incorrect judgement as before, but also judge the extent of correctness.

In Fuzzy Set Theory, a *set* is defined as $A=\{a_1, a_2, a_3, \ldots, a_n\}$, where A means *universe of discourse*, and $a_1, a_2, a_3, \ldots, a_n$ are all *members* or *elements* of A. When every member of set A is also a member of set B, we call A a *subset* of B, denoted by $A \subseteq B$ (Partee, Meulen & Wall 2009).

For a given universe of discourse A, the mapping from A to unit interval [0, 1] is called a *fuzzy set* of A, written as μ_A: $A \rightarrow [0,1]$. A fuzzy set is an ordered set, meaning that the order of the members is unchangeable.

In order to measure the similarity or distance between the two fuzzy sets, an index termed *Single Similar Degree* is used, which can be defined as follows: Suppose $U=\{u_1, u_2, u_3, \ldots, u_n\}$, $A, B \in P(U)$, and we define an operator δ, then if the mapping $\delta: P(U) \times P(U) \rightarrow [0,1]$ satisfies $\delta(A, A)=1$, $\delta(B, B)=1$, and $\delta(A, B) \geq \delta(A, C)$ if $A \subseteq B \subseteq C$ or $C \subseteq B \subseteq A$, then $\delta(A, B)$ is the *Single Similar Degree* of A to B. $\delta(A, B)=m/n$, where m is the number of the members in A appearing in B, and n the number of the members in A.

From the above definition, it can be inferred that Single Similar Degree does not satisfy the property of commutativity, i. e. that the change in the order of the members in an operation leads to a changed operation answer. Here, $\delta(A, B)$ and $\delta(B, A)$ have different meanings and algorithms. $\delta(A, B)$ is the Single Similar Degree of A to B, while $\delta(B, A)$ is the Single Similar Degree of B to A.

In the present study, Single Similar Degree is used to measure the similarity between the reference answer and the student's response. In English, a sentence made up of some words in a certain order can be regarded as a fuzzy set $U=\{u_1, u_2, u_3, \ldots, u_n\}$. Suppose Sentence A is the reference answer and Sentence B is a student's response, then $P(A)$ and $P(B)$ are two fuzzy sets representing the two sentences respectively. Supposing the number of words in $P(A)$ is n, and the number of words in $P(A)$ appearing in $P(B)$ is m, then $\delta(A, B)=m/n$, indicating the Single Similar Degree of the reference answer to the student's response.

The Single Similar Degree of the reference answer to the student response is labelled Single Similar Degree I. Since there may be many reference answers, they will be considered as many fuzzy sets. In order to find which reference answer is most similar to the student response, we need to calculate the Single Similar Degrees of various reference answers to the student response respectively. The reference answer with the highest Single Similar Degree I will be regarded as the most similar one to the student response, which then will be assigned the corresponding score attached to that reference answer. However, sometimes two or more reference answers may lead to the same highest Single Similar Degree I. Therefore, conversely, we need to calculate the Single Similar Degree of the student response to those two or more reference answers respectively, which is labelled as Single Similar Degree II. The score of the reference answer which is in accordance with the highest Single Similar Degree II will be assigned to the student response. From this, it can be seen that in the present study, Single Similar Degree I is the primary marking index, while Single Similar Degree II is the secondary marking index.

By definition, a word is 'the smallest of the linguistic units which can occur on its own in speech or writing' (Richards, Platt & Platt 2000:510). In writing, word boundaries are usually recognised by spaces between words. So it is easy for computers to identify any English word. Apart from the obvious, a computer should count any number (e.g. 1950s, 187200, 6:30), abbreviation (e.g. U.S.A., U.N.E.S.C.O., p.m.) or compound word which contains a hyphen or hyphens (e.g. Anglo-Saxon, first-hand, 5-year-old) as a single word. In this study, such a condition may exist: a long word in one fuzzy set contains a short word in another fuzzy set (e.g. *however* and *ever*). Under such circumstances, a computer may 'consider' that the two sets have the same word (*ever*). In order to solve the problem, we add one space before and after any word.

What follows is an example of how Single Similar Degree is calculated. In this example, the sentence *He is a very good and handsome boy* is the correct answer which carries 2 points. For the purposes of auto-marking, this sentence is a fuzzy set $S_1=\{He, is, a, very, good, and, handsome, boy\}$. If a student's response was *He was a good but ugly boy*, this response would constitute another fuzzy set $S_2=\{He, was, a, good, but, ugly, boy\}$. The computer finds all the words in S_1 which appear in S_2. The total number of S_1 words which appear in S_2 is $m=4$. The total number of the words in S_1 is $n=8$. Therefore, the Single Similar Degree $\delta 1(S_1, S_2)=m/n=4/8=0.5$. Similarly, as the total number of words in S_2 is $n_0=7$, the Single Similar Degree $\delta 2(S_2, S_1)=m/n_0=4/7=0.57$.

Methodology

On the basis of the Theory of Single Similar Degree and its algorithm, an auto-marking system on SAQs was designed and then improved through three experiments.

Participants

The participants in this study comprised 220 non-English majors. They were randomly chosen from a sample

university and invited to take a test on SAQs. Test items were from the test paper of College English Test Band 4 (CET-4, a national English test for undergraduates in China) that was administered in June 2005.

Two human raters participated in the study. They were English teachers at a university, and had received professional training in CET-4 SAQ marking.

Data collection

Two types of data were collected in the study: 1) the marks awarded by the two human raters and 2) the marks obtained through the auto-marking system. The process of data collection consisted of several steps. Firstly, the two human raters marked each part of the test papers respectively. The first rater's mark (R_1M) and the second rater's mark (R_2M) were compared. If the two markers gave one response different marks, the average of R_1M and R_2M was taken as the final mark (FM). Secondly, the auto-marking system was used to mark the test papers and get the mark obtained using the system (SM). Thirdly, R_1M, R_2M, SM and FM were compared to calculate their marking correctness ratios against the total items. Finally, SPSS version 16.0 was used to carry out a correlation analysis on the marks of R_1M, R_2M, SM and FM.

Findings and discussion

The first experiment

The system adopted in this experiment is based exclusively on the Theory of Single Similar Degree. In the experiment, the system was used to mark 220 test papers; each part of the papers included eight test items, making a total of 1,760 test items.

Drawing on the data presented in Tables 1 and 2, the findings are summarised below:

1. The marking correctness ratios of R_1 and R_2 are very close to each other (see Table 1) and the correlation between R_1M and R_2M reaches .954 (see Table 2). They indicate that inter-rater reliability is very high.
2. Although the two human raters have received formal training of CET-4 SAQs marking and their marking correctness ratios reach 92.22% and 91.02% respectively, these ratios are still much lower than those of the most advanced SAQ marking system – Automark whose correctness ratio has reached 93–97%. One of the reasons for the lower correctness ratio achieved by the human raters may be that the raters' subjectivity affects their judgment in marking. There may also be many other factors which are hard to control in the human marking of subjective questions.
3. The marking of the system in this application cannot be deemed satisfactory. Out of the total 1,760 test items/available marks, 350 marks obtained by the auto-marking system were incorrect. Its correctness ratio achieved only 80.11%, which is considerably lower than that of R_1 and R_2 (92.22% and 91.02%) and far below the level achieved by Automark. This made it clear to the researchers that the system needed to be improved.

Table 1: Marking results

Marking instrument	Total items	Correctly marked items	Marking correctness ratio
R_1*	1,760	1,623	92.22%
R_2*	1,760	1,602	91.02%
S*	1,760	1,410	80.11%

* R_1: the first human rater, R_2: the second human rater, S: the auto-marking system on SAQs

Table 2: Correlations between the four sets of marks

	FM*	R_1M	R_2M	SM
FM	1.000			
R_1M	.957**	1.000		
R_2M	.952**	.954**	1.000	
SM	.768**	.730**	.721**	1.000

* FM = final mark, R_1M = first rater's mark, R_2M = second rater's mark, SM = auto-marking system mark
** Correlation significant at the 0.01 level (2-tailed)

All of the 350 wrongly marked items were inspected. It was found that some of the frequent responses to those items were not anticipated and had not, therefore, been listed among the reference answers in the original mark scheme for SAQs. For example, many responses to Item 8 of the SAQ, *What should a sportsman do to avoid killing a rare species of wildlife?* (1 point), were listed in the reference answers (see the Appendix). However, one response *'be sure of the identity of the target'* was not included, which was the response of 60 students to this item. Although it is easy for human raters to find this out-of-the-list answer acceptable, the auto-marking system assigns 0 points to it, for it is most similar to the reference answer '*Identity of the target (0 points)*', based on Single Similar Degree I.

Unlike human raters, computers cannot deal with fuzzy problems, such as the extent to which a response is acceptable. This is one of the greatest limitations of the auto-marking system. These problems are termed 'fuzzy' because they cannot be described clearly in a mathematical way. Only the aspects of the natural language which can be transformed into numerals can be dealt with by computer. In view of this and based on the suggestions of Bachman & Palmer (1996) on the scoring of limited production tasks, test developers need to develop a scoring key listing a wide range of responses that could be considered acceptable to some degree. Making the scoring key as comprehensive as possible should help improve the marking correctness ratio.

The responses which appeared frequently in students' test papers (regardless of whether they were correct or incorrect as judged by human raters), but which were not already included in the reference answers, were summarised and added to the reference answers with their corresponding credits. The new version of the reference answers was adopted as the basis for the system to be tested in two experiments which are discussed next.

The second experiment

All conditions in the second experiment were the same as those in the first experiment, except that this application was

based on the new version of reference answers. The marking results are presented in Tables 3 and 4.

Table 3: Marking results

Marking instrument	Total items	Correctly marked items	Marking correctness ratio
R_1	1,760	1,623	92.22%
R_2	1,760	1,602	91.02%
S	1,760	1,503	85.40%

Table 4: Correlations between the four sets of marks

	FM	R_1M	R_2M	SM
FM	1.000			
R_1M	.957**	1.000		
R_2M	.952**	.954**	1.000	
SM	.769**	.726**	.721**	1.000

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

From Tables 3 and 4, it can be seen that the marking result for the system in the second experiment is better than it was in the first experiment. The marking correctness ratio rises from 80.11% to 85.40%. This shows the importance of specifying a wide range of acceptable reference answers. Nevertheless, the marking correctness ratio for the system is still well below that of the human raters.

From a further analysis of the test items, the marks and their Single Similar Degrees, it was found that nearly all the students' responses with Single Similar Degree I below 0.6 should have been assigned 0 points rather than 0.5 points or higher. Table 5 illustrates some of the students' responses which were marked wrongly by the system.

From Table 5 it is clear that these students' responses are not even close in meaning to the reference answers and should be assigned 0 points. However, the system judged that these students' responses are still a little similar to the reference answers since they still contain one or a few identical words. However, the Single Similar Degree I of these responses is very low; the highest is no more than 0.6. Out of 513 student responses, 510 should have scored 0. Out of 254 marking errors made by the system in the experiment, 183 errors were caused by the system not marking a response zero. The wide existence of such a phenomenon in the marking process greatly influences the scoring accuracy of the auto-marking system. It indicates that there must be an underlying rule affecting the correctness of marking.

Consequently, a careful analysis of the 183 errors mentioned above was conducted. Two points emerged. Firstly, these students' responses were not really similar to the reference answers since their Single Similar Degrees were very low, at most 0.6. Secondly, students' responses and reference answers were only identical in unimportant words. Here, 'unimportant words' are defined as the words which do not carry the content of the response and which are not key words. According to Hu (2001: 80), content words 'carry the main content of a language referring to substance, action and quality'. Content words are usually nouns, verbs, adjectives, and adverbs. Thus, a marking index concerning content words should be added to the system. In the study, 'key word' is used rather than 'content word' in the context since it is used more commonly and can be understood easily. In the third application of the system, Key Word Ratio was added as a marking index, as discussed below.

The third experiment

In this experiment, the index of Key Word Ratio was added to the system. The Key Word Ratio algorithm consists of the following: firstly, list the key words of a reference answer and suppose the number of these key words is *A*; secondly, calculate how many of these key words have appeared in the student response, and label their number *B*. Thus the Key Word Ratio is *B/A*. Since there may be many reference answers for one item, any student's response may have many key word ratios. It was decided to use the largest Key Word Ratio as an important marking index which will influence the score of a response. If the largest Key Word Ratio of an answer is less than 0.6, zero points will be assigned to the answer and the index Single Similar Degree I and Single Similar Degree II will be adopted.

A student's response to test item 7 was selected as an example:

S7: *What are people advised to do before they remove illegal or undersized fish from the hook?*

Wet their hands. (2 points)

Table 6 displays a student's response to test item 7 and several reference answers.

Table 5: Database of students' scores (an extract)

Students' response	Reference answer	Single Similar Degree I	Single Similar Degree II	Score of reference answers
Don't be a game hog	Fish and game laws	0.25	0.20	1
To be punished as a violator	Don't be a violator against the sports law	0.38	0.50	1
The right reason	Moving in the right way	0.40	0.67	0.5
Be sure of the identity of your target before you shoot	Wet the hand	0.33	0.09	2
Don't use the gun	Identifying the target	0.33	0.25	0.5
Thoughtless and to kill whatever flies within range	To be sure of the identity of target	0.13	0.13	0.5

Table 6: A student response and its related information

Student's response	Reference answer	Key words	Key Word Ratio	Single Similar Degree I	Score of reference answers
Be sure of the identity of your target before you shoot	They should wet their hands	wet, hands	0	0.00	2
	Wet their hands	wet, hands	0	0.00	2
	To wetting hands	wetting, hands	0	0.00	1.5
	Wet their hand	wet, hand	0	0.00	1.5
	This should be done only after wetting the hands	Done, wetting, hands	0	0.22	1
	They are advised to do this only after wetting the hands	advised, do, wetting, hands	0	0.09	1
	After wetting the hands well	wetting, hands	0	0.20	0
	Take a knife, cut the line or leader as close to the hook as convenient	take, knife, cut, line, leader, close, hook, convenient	0	0.07	0

From Table 6, it can be seen that the student's response is not even partially correct: its meaning has nothing to do with the reference answers. According to the algorithm of Key Word Ratio, its Key Word Ratio is 0 and it should indeed be assigned 0 points. However, if it is marked according to the algorithm of Single Similar Degree I, the student should be assigned 1 point. Consequently, the importance of Key Word Ratio becomes clear: it is more effective to mark the content of a student's response by incorporating the algorithm of Key Word Ratio into the auto-marking system than to use Single Similar Degree. This is illustrated by the results of the third experiment reported in Tables 7 and 8 below.

Table 7: Marking results

Marking instrument	Total item no.	Correctly marked item no.	Marking correctness ratio
R_1	1,760	1,623	92.22%
R_2	1,760	1,602	91.02%
S	1,760	1,614	91.07%

Table 8: Correlations between several marks

	FM	R_1M	R_2M	SM
FM	1.000			
R_1M	.957**	1.000		
R_2M	.952**	.954**	1.000	
SM	.865**	.826**	.826**	1.000

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

Tables 7 and 8 indicate that the marking correctness ratio of the system has risen to 91.07%, which is very close to what was achieved by the two human raters in this study. Besides, the correlations between SM and R_1M, R_2M, and FM are all over .80. This indicates that Key Word Ratio is crucial as it marks the meaning of a student's response more accurately than the other two indexes.

Conclusions and limitations

The present study discusses an auto-marking system for SAQs which was designed on the basis of the Theory of Single Similar Degree in Fuzzy Mathematics and which was gradually improved through three experiments. The following conclusions are drawn from the discussion of the results of these experiments:

1. Single Similar Degree plays a very important role in measuring the similarity between the student's response and reference answers. Based exclusively on Single Similar Degree, the marking correctness ratio of the system can reach 85.40%.
2. Single Similar Degree does not function quite so well when used on its own; Key Word Ratio was found to be very important in marking the semantic content of a student's response. Consequently, Key Word Ratio was added to the system, which increased the marking correctness ratio of the system to 91.07%. This is 5.67% higher than the ratio achieved by the system when only Single Similar Degree was used.
3. It was also suggested in the study that the acceptable reference answers for auto-marking should be as wide-ranging as possible. This is necessary in order to avoid the auto-marking system 'rejecting' a response which would have been found acceptable by human raters.

Although the marking correctness ratio of the system has reached 91.07%, which is close to that of the human raters who participated in this study, the experiment scale is limited and the design is still at the initial stage. Therefore, the system cannot be applied on a large scale yet. Despite this, the study has explored certain factors which affect the marking accuracy of the auto-marking system, and could offer some reference points for future research studies.

References and further reading

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

Burstein, J, Kaplan, R, Wolff, S and Lu, C (1996) *Using lexical semantic techniques to classify free responses*, paper presented at the Proceedings of SIGLEX 1996 Workshop, Annual Meeting of the Association of Computational Linguistics, University of California, Santa Cruz.

Gao, S and Yuan, C (2004) The Application of Sentence Similarity Measurement in Automated Assessment Technology of Subjective Tests, *Computer Engineering and Applications* 5, 132–134.

Hu, Z (2001) *Linguistics: A Course Book* (2nd Ed.), Beijing: Peking University Press.

Liang, M (2005) *The Construction of an Automated Scoring System on Chinese Students' English Essay*, unpublished PhD dissertation, Nanjing University.

Liang, M and Wen, Q (2007) A Critical Review and Implications of Some Automated Essay Scoring Systems, *Computer-Assisted Foreign Language Education* 5, 18–24.

Meng, A, Bu, S, Li, Y and Gan, W (2005) Design and Implementation of an Automated Assessment Algorithm for Subjective Tests in Network Examination System, *Computer & Digital Engineering* 7, 147–150.

Newble, D and Cannon, R (1989) *A Handbook for Teachers in University and College*, New York: Kogan Page.

Partee, B H, Meulen, A and Wall, R E (2009) *Mathematical Methods in Linguistics*, Beijing: Beijing World Publishing Corporation.

Richards, J, Platt, J and Platt, H (2000) *Longman Dictionary of Language Teaching and Applied Linguistics*, Beijing: Foreign Language Teaching and Research Press.

Valenti, S, Neri, F and Cucchiarelli, A (2003) An Overview of Current Research on Automated Essay Grading, *Journal of Information Technology Education* 2, 319–330.

Zadeh, L (1965) Fuzzy Sets, *Information and Control* 8, 338–353.

Appendix

Reference answers to Item 8:

(1 point)

Identify the target.

Be sure of the identity of their targets.

Be/Make sure of the identity of the target before shooting.

Identify their target/targets.

They should know the identity of the target.

(0.5 points)

To be sure of the identity of ^ target.

Identify his/your target.

Make sure of the species of ^ target.

Identifying the target.

Being sure of the identity of the target.

Make sure ^ the identity of the target.

Making sure ^ the identity of the target before shoot.

(0 points)

Insure the identity of target.

Identity of the target.

Teacher learning on the Delta

SIMON BORG SCHOOL OF EDUCATION, UNIVERSITY OF LEEDS, UK

Introduction

Teacher learning on the Delta was a project funded by Cambridge ESOL to examine the learning experienced by candidates on the *Diploma in Teaching English to Speakers of Other Languages* (Delta) teaching qualification and to assess the impact of this learning on their professional practices subsequent to the course. This article provides an overview of this project and its findings.

Background

The context for the project was the Delta, an internationally recognised advanced course for practising teachers[1]. This qualification can be taken in a modular fashion (there are three modules which can be taken separately) or as a full-time integrated course. In the latter case, which was the context for this study, candidates teach 10 lessons to adults (five are observed and assessed). They are also expected to spend a substantial number of hours on reading, research and assignment writing. Significantly revised in 2008, the Delta seeks to reflect contemporary views of good practice in language teacher education (see Zeronis 2007 for a discussion of the development and design of the course). The full-time version of the Delta was the focus of this study.

In commissioning this project, Cambridge ESOL was interested in understanding the ways in which the Delta impacted on teacher learning, both during the course and when candidates returned to work. There have been a number of studies of the impact of language teacher education in pre-service contexts (e.g. Borg 2005, Busch 2010), and work also exists in education generally about the impact of continuing professional development on teachers (e.g. Goodall, Day, Lindsay, Muijs & Harris 2005). Research into the impact of in-service language teacher education is, however, scarce. Lamie (2004), for example, reported the positive impact of in-service training on developing more communicative orientations to language teaching in four teachers in Japan, while Freeman (1993) and Scott & Rodgers (1995) also provide evidence of ways in which language teachers' beliefs, attitudes and understandings changed through in-service teacher education. The papers in

[1] See <http://www.cambridgeesol.org/exams/teaching-awards/delta.html> for full details of the course.

Hayes (1997) present further perspectives on the impact of in-service language teacher education.

Of particular relevance here is the work of Phipps (2007, 2010), who examined the impact of a Delta course on one teacher, though the format of the programme he studied was very different to that studied here (it was integrated into an MA programme and taught part-time over 18 months). Phipps (2007) focused on the initial four-month phase of the course, and while he found that 'there were few tangible changes to existing beliefs' about grammar teaching, there was evidence of developments in the teacher's awareness and strengthening of their beliefs (ibid:13). Similarly, while there were no radical changes to the teacher's classroom practice, the Delta did impact on the depth of planning and thinking the teacher engaged in. This study also highlighted developments in the teacher's confidence. These conclusions were reinforced in his more extended analysis of change in three teachers' beliefs over the full 18-month period of the Delta (Phipps 2010). In contrast, Lamb (1995) found, one year after running an in-service course for language teachers in Indonesia, little evidence that the course had impacted on their beliefs, while Kubanyiova (2006) also reported on the limited impact of an in-service teacher development programme on a group of English language teachers in Slovakia, concluding there was 'no indication of whether or not change in the teachers' beliefs took place' (ibid:7). While Richards, Gallo & Renandya (2001) found through their survey that teachers said that in-service training was a major influence on changes in their practice, they did not provide specific insights into the nature of such impacts. Overall, then, the *Teacher learning on the Delta* project, while of immediate relevance to the work of Cambridge ESOL, was also well-positioned to make an important contribution to the language teacher education literature.

Methodology

Research questions

The study addressed the following research questions:

1. What do participants expect to learn on the Delta?
2. What impact on their professional practice do participants expect the Delta to have?
3. What evidence is there of teacher learning – defined broadly as changes in knowledge, practices, awareness, attitudes, and beliefs – both in participants' own accounts of their experience while taking the Delta modules and in their assessed work and tutor feedback on it?
4. Do the candidates feel that certain learning activities and processes on the Delta make a particular contribution to teacher learning?

After completing the Delta, what impact on their professional practice do participants say the course has had?

Context and participants

The Delta course studied here was taught at a training centre in the UK full-time (i.e. all day, five days a week) over an eight-week period. Six Delta candidates agreed to take part in this study (out of a total of 12 who were registered on the course when volunteers were sought via a preliminary questionnaire). All six teachers were female, British and worked in private language teaching organisations. They had between two and 10 years' experience in English language teaching (ELT) and held a range of positions from teacher to Director of Studies in the schools they worked for.

Data collection and analysis

Guskey (2000) suggests that studies of the impact of professional development activities are often limited because they rely on questionnaire data, occur as one-off activities, and are typically summative. The approach to examining the impact of the Delta taken here addresses these concerns by being longitudinal and drawing on a range of qualitative data. A qualitative perspective allowed for the detailed analysis of participants' experiences of the Delta, while the longitudinal dimension made it possible to track teacher learning and its impact during and beyond the course. Table 1 summarises the data collected for each teacher. Collectively, this data provided a substantial database of some 100,000 words per participant which provided evidence of their work on the Delta, assessor feedback on it, and participants' reflections on the impact of the course on them while studying and on returning to work.

This data was subjected to an analytical process which was wholly qualitative (see, for example, Bryman 2008 for an overview of key issues in working with qualitative data). The analysis was:

- cyclical (analysis took place in between, and informed, each subsequent phase of data collection)
- iterative (the data for each teacher was worked through several times).

The analysis involved:

- progressive focusing – i.e. moving from a large volume of data covering several themes for each candidate to a more concise analysis of key specific themes of relevance to each case

Table 1: Data collected for each teacher

• Pre-course interview tasks
• Preliminary questionnaires
• Six audio recorded interviews (two in person and four by telephone, distributed across the course)
• All coursework submitted for the three Delta modules – the diagnostic assignment, four Language/Skills assignments, the Exploratory Practice assignment, and the three Reflection and Action assignments
• The Extended Assignment
• Feedback provided by tutors and external examiners on all coursework
• Written feedback from the teachers on their case study reports, including comments on the impact of the Delta six months after its completion.

- the inductive generation of analytical categories (so that these emerged from the data)
- extensive use of contextualised extracts of primary data in the emerging reports (e.g. quotations from the teachers and their work)
- respondent validation (through which teachers were asked to comment on the extent to which my accounts of their experience on the Delta were ones they could relate to).

Through these processes key themes in the teachers' experiences on the Delta were identified, and data relevant to these themes marshalled in order to create narratives for each teacher. The final accounts (each some 10,000 words long) were thus representations, chronologically arranged, of the development each teacher experienced on the Delta and the impact that the course had on them. The teachers were invited to read their accounts and to provide feedback on two issues: (1) the extent to which they felt the account was an accurate representation of their experience on the Delta and (2) the extent to which the Delta had continued to impact on their work six months down the line. Five teachers provided feedback, all confirming that they were happy with the accounts. For example T1 wrote that 'I think it is an excellent account of my Delta experience … I would say this is definitely a story I can relate to' while T5 replied that 'Yes, this was a very accurate account of my experience'[2].

In terms of key ethical concerns in educational research (see, for example, Denscombe 2002), voluntary informed consent was obtained from all participants, they had the right to withdraw from the study at any time, and their anonymity was protected together with the confidentiality of the data. Participation in the study was incentivised (the teachers received a book token at the end of the project). This was to acknowledge the willingness of the teachers to take part in this project during eight very intensive weeks of study on the Delta.

Findings

I will now summarise key findings in relation to the research questions listed above. It is not my goal to explore any particular issues in great depth here (this will be the focus of separate publications) but to provide an overview of the issues that emerged from the study.

Expectations of learning and impact

In relation to research questions 1 and 2, participants were asked early in the study what they expected to learn on the Delta and what difference they felt it would make to their subsequent practice. Their responses indicated an interest in both theoretical and practical knowledge, with the theory of ELT in particular often being singled out as the area participants most hoped to learn about. There was also a general concern among the participants for becoming revitalised, more confident professionals with a stronger rationale for their work. One teacher, for example, said that 'recently I have become aware of a slight stagnation in my teaching methods and I would like to improve the quality and variety of my teaching approaches' (T3:R&A2). Participants with responsibilities for supporting teachers in their institutions also expressed an interest in enhancing their ability to fulfil that dimension of their role.

Learning activities and processes

In relation to research question 4, at various stages in the study participants were asked about particular activities, processes and mechanisms on the Delta that they felt facilitated learning. An overall finding here is that there was variation in their evaluation processes that were designed to support teacher learning both across participants as well as within the experiences of individual participants at different points of the course (e.g. not all teachers saw value in reflective writing). The processes listed in Table 2, though, were identified by most participants at some point on the Delta as being beneficial.

Table 2: Valued learning activities and processes on the Delta

Activity/Process	Illustrative quotation
Reading	'It gives you more confidence in the sense that you know why you're doing certain things. It gives you ideas because it suggests … approaches and ways of doing things that perhaps you were doing but you didn't really follow it all the way through' (T2:I2).
Peer feedback on teaching	'I think they've been really good actually … it's quite nice to have an outsider's look at things, but also an experienced outsider, because they're all fellow teachers. So it's been quite good to get them to look at these different things that I wanted to concentrate on' (T6:I3).
Observing peers	'You see lots of different styles of teaching and different personalities and strengths and weaknesses of everybody's style. So it's quite useful too – I think you do learn a lot always from watching other people' (T5:I3).
Tutor input sessions	'They're brilliant. Sometimes demo lessons, sometimes lectures almost, it's not a lecture because you can ask questions and interrupt to a certain point … and sometimes quite practical, so they're, nice variety, and, yeah, really interesting, I think' (T6:I2).
Tutor feedback on teaching	'I really believe it is the feedback that I have received from my tutors on the course that I have found most valuable as it is always pertinent, relevant and thought-provoking. Without their feedback, I would have been unable to reflect on my experiences and beliefs as successfully … it is undoubtedly what has helped me develop the most' (T4:R&A4).
Experimental Practice*	'I feel this experiment was an extremely effective way of highlighting my unnecessary verbal communication. Gestures and longer pauses, which gave students time to reflect and offer their own/peer corrections, dramatically reduced my verbal input' (T3:EP R&E).

* This is an assignment in which candidates conduct a teaching experiment using a particular instructional approach, strategy or framework – e.g. task-based learning or the silent way.

[2] Data cited in this paper carries the following codes: T1, T2 etc. = individual teachers; I1, I2 etc = first, second etc. interview; FB = tutor feedback; R&A2-4 = reflective assignments; CSFB = teacher feedback on their narratives; PCT=pre-course task; EP=experimental practice; R&E=Post-lesson self-evaluation; BE=background essay.

Teacher learning during the course

In this study, teacher learning during the Delta was assessed in a range of ways: (a) examining teachers' language and skills assignments, (b) studying their professional development portfolios, and (c) asking teachers to reflect on their learning through the interviews during the course. I summarise the findings for each of these areas in turn below.

Language and skills assignments

For each of the topics focused on for their language and skills assignments, participants had to examine relevant literature and write a background essay meeting specific academic writing criteria. They also had to examine practical teaching problems relevant to these topics and to identify solutions to these problems. Finally, they had to produce a detailed lesson plan informed by their prior background essay and to teach this lesson. The depth of analysis required in these assignments (as well as the fact that the topics were new to the participants) meant that participants inevitably expanded their understandings of the chosen topics significantly. Thus, T3 concluded her LSA1 background essay on collocation by noting the following:

> Having examined the form, meaning, use and pronunciation of lexical collocations in parts of speech, as well as examining ways of overcoming some of the most pressing problems for both the learner and teacher, I feel much better equipped to help learners explore such collocations (T3:LSA1 BE).

T4, to take another example, concluded her LSA3 essay on accuracy in spoken English by saying that 'I have gained a more in-depth awareness of the different kinds of accuracy in spoken English, and how the features of speech relate to the skills students need to perform accurately' (T4:LSA3 BE).

Even where participants focused on areas they felt they already knew about, by the end of the process they had typically reviewed this initial assessment and acknowledged that perhaps their prior knowledge was not as solid as they had assumed. Reflecting on her LSA3 work on extensive reading, for example, T2 noted that 'I thought I knew quite a lot and in fact I don't know anything … When I started reading about it, I realised just how little I did know' (T2:I4).

The assignments also provided participants with opportunities to extend their understandings of ideas encountered through reading and input sessions. For example, early on in the course the teachers were introduced to the concepts of guided discovery and restructuring in relation to grammar teaching, and reference to these concepts was frequent, especially in early assignments. In her assignment on 'get' as a delexicalised verb, T1 thus explained:

> I have decided that in order for them to unpack the meaning and form of the target language I will use the guided discovery technique. Without trying to terrify the learners I aim to encourage them to see some of the different combinations that it can be found in. In consciously raising their awareness of collocations and multi word verbs I hope that their mental lexicon will restructure in order to make way for a part of lexis that occurs highly frequently (T1:LSA1 LP).

Given that teachers were producing work for assessment, there was of course an element of knowledge display in their assignments. Participants could not, however, get away with simply reciting theoretical knowledge in their background essays; they were also required to demonstrate a sound grasp of its practical applications by designing and teaching a related lesson. Clearly, then, participants enhanced both their theoretical and practical knowledge of the topics they focused on for their assignments, and this represented a significant dimension of teacher learning on the Delta.

Professional Development Assignments

A second way of examining teacher learning on the Delta is to examine the themes participants focused on in their Professional Development Assignment (PDA). Individual teachers followed unique development pathways during the course – i.e. while they worked within the common framework provided by the Delta, the particular issues that became the focus for their development during the course varied. An analysis of teachers' PDAs highlights the origins of the developmental foci that each pursued (most commonly these emerged from feedback on observed teaching) and teachers' attempts to address them through cycles of teaching, feedback and reflection. In most cases significant breakthroughs were evident, while some issues remained as areas for teachers' continuing development after the course.

Table 3 presents two examples of the kinds of teacher learning evidenced in the PDAs. T1 identified over-planning as an area for development early in the programme and it remained a focus throughout her PDA; by her final Reflection and Action (R&A) entry, she had reached a deeper understanding of why she over-planned and was able to make sense of this in relation to her lack of confidence in teaching grammar. Such awareness is a form of teacher learning. In the second example, there is evidence in T6's PDA of development in her understanding of lesson shapes – particularly in relation to the timing of the 'production' stage; whereas she had always seen this as a terminal activity,

Table 3: Development in Professional Development Assignments

Teacher	Focus	Statement of problem	Example of teacher learning
T1	Reducing over-planning; becoming more confident in her knowledge of grammar.	'I am very thorough in my planning … Unfortunately this can lead to over-planning which often leaves me feeling like I haven't achieved my end goal' (T1:R&A2).	'I have low confidence when it comes to my linguistic knowledge and I'm scared of not being able to answer the learners' questions. This leads me to over-plan and over complicate the lesson with extra activities' (T1: R&A4).
T6	Providing longer and earlier production stages.	'I am not offering [the] production stage early enough in the lesson or for long enough' (T6:R&A3).	'before this course, I always thought, 'oh it [production] comes at the end' … but actually what I've realised is that that turns on its head and the language analysis can come after and that gives them [the learners] a chance to improvise and try it out first … and I think I've realised that that works much better … the shape of my lesson has completely changed because of that' (T6:I4).

Table 4: Impact of Delta on teacher learning

Areas of impact	Illustrative quotation
Confidence to talk theoretically about teaching	'I'm much more confident that … I could successfully have a debate with anyone … about teaching theories and stuff like that whereas I wouldn't even have bothered before' (T3:I6).
Greater attention to rationale for practices	'I think before I was so focused on what I'm going to do each day … that I didn't spend time thinking about why I'm doing things and whether this is actually the best way to do things and I think the impact that the course will have, definitely when I go back to work, is that I think I will think about those things a lot' (T5:I5).
Enhanced range of practical techniques	'I think we have learnt lots … some things that were new and some things that I did know but hadn't really been doing … so yeah, definitely I've learnt actual different ways to approach tasks as well … the biggest thing I've learnt about is being more student-centred so I'll probably try and make my activities more student-centred, more investigative work and discussion and less teacher-led' (T5:I5).

through the course she had been encouraged to consider alternatives. Her comments in I4 demonstrate changes in her thinking on this matter.

Teachers' reflections on their learning

Teachers were interviewed periodically during the programme and asked about their learning; in particular, the interview that took place just after the end of the Delta (I5) engaged teachers in considering the impact they felt it had had on them. An analysis of their comments provides further insight into the impact of the Delta and showed that the teachers felt this impact had been significant. They said it had developed their theoretical and practical knowledge of ELT, enhanced their awareness of their beliefs, and boosted their confidence in their abilities as teachers and advisors of other teachers. They also felt they were better able to justify their own practices, more critical in their questioning of their practices, and better able to talk about teaching. It is clear, then, that participants felt the Delta had impacted on their learning significantly, though there was of course variability in the range of development each candidate experienced in relation to particular issues (e.g. not all teachers agreed that the Delta led to any significant changes in their beliefs about language teaching and learning – see Borg (forthcoming) for a specific analysis of this issue).

Table 4 provides examples of how the teachers talked about the impact of the Delta at the end of the course.

Post-course impact

The final objective of this study was to assess the impact of the Delta on participants on their return to work. Evidence relevant to this issue was obtained via the final interview in this project, which was conducted two months after the end of the course. Additional evidence was provided six months after completion of the course when participants provided feedback on their cases studies.

Overall, participants were very positive about the impact that the Delta had had on their professional practice since returning to work. They highlighted enhancements in relation to the following range of areas of their work as ELT professionals:

- confidence in observing and supporting other teachers
- reflective skills
- ability to participate more fully in discussions about language teaching at work
- confidence in the classroom
- ability to evaluate courses
- ability to work as a teacher trainer
- credibility as a manager
- the sense of value they saw in their work as teachers
- a sense of control over their own practices
- speed and focus of planning
- ability to improvise
- ability to implement student-centred practices
- awareness of the rationale for their pedagogical choices
- broader views of the role of the teacher.

One teacher, for example, explained that after the Delta:

> [I] came back to work and just believing in myself, being much stronger about what I do in the classroom, and just being able to go with students and when they come back with things say, right ok, let's work on that, let's expand that. So it does, it gives you more confidence (T2:I5).

Another explained that 'I feel like I have a lot more control over what I do from a theoretical point of view, like I know why I'm doing things or why I'm making these decisions' (T3:I6). This deeper understanding of what they do also allowed teachers to be more spontaneous, as noted by the teacher who said that 'I feel freer to make decisions on the spot, to adapt plans and change things. I feel freer to be flexible, because I know that I can justify what I'm doing' (T4:I5).

In addition to these points, it should be noted that five of the six participants were promoted soon after the Delta and thus career progression was another very immediate form of impact the course had for these individuals.

One particular aspect of post-course impact which merits a comment here is participants' classroom practices. While change in these was reported, two participants were also of the view that there was not much observable change in their classroom behaviours after the course compared to before. One participant noted that:

> I don't think I'm a particularly different person in the classroom. So practically, I think I've learnt a couple of things … things to try and not do so much of, things I've tried to carry on doing well. But I don't think I have acquired a huge amount of practical skills (T6:I5).

Another felt that 'I suppose in terms of whether my students will actually notice a difference, I hope they will and

there's a slight difference but generally the day to day life continues fairly consistently' (T3:I6).

What these participants stressed had changed was the thinking behind their work, their awareness of their learners and of the rationale for their pedagogical decisions, and their interactive decision making. This is an important observation about in-service teacher education generally; with experienced teachers, the impact of teacher education may lie less in dramatic changes in what they do than in the strengthening of their understanding of their established practices. This is a reminder that evidence of the most significant impacts of teacher education may not always be directly observable; this may be particularly true, where the participants are experienced practitioners who are generally satisfied with themselves as teachers and are seeking not to radically overhaul what they do but rather to develop a theoretical rationale for their existing practices and to make some adjustments and refinements to these. Thus while some teachers in this study did report considerable changes in their classroom practices, it is also clear here that significant teacher learning can occur without major observable changes in what teachers do.

In discussing the Delta's subsequent impact on the teachers, it must also be noted that this was inevitably shaped by contextual factors in their institutions. T3, for example, had to follow a coursebook that the students purchased and was required to work through it at a certain pace. T4, in describing a specialist business English course she was teaching on returning to work, explained that it was not the kind of course where she could experiment, pedagogically. And T6, together with T3, both noted that they would not have time to plan lessons in the detail required on the course. T5 also noted that the climate of her school made it difficult for her to feel part of a professional community there and hence to engage in the kinds of collaborative development activities she had hoped to pursue after the course. Though these contextual constraints were not a major theme in the data for this study, they must nonetheless be borne in mind in assessing the extent to which the Delta shaped participants' practices when they returned to work.

At the end of the Delta, teachers were unanimously very positive about the experience. One illustrative quote here represents the sense of achievement with which all teachers reflected on the course:

> Massive. It's been, I suppose it's a bit silly to say life changing, but it's been a complete revelation. I feel I've got an awful lot out of it in all sorts of areas … I didn't expect to like the theory as much as I did, just so many things that now that I'm back at work, I'm looking through things and thinking, 'oh yes, of course, yes, no, we did this and we studied that, I know why we're doing that and, oh no, I don't agree with that'. It's, yeah, it's been a revelation, it really has (T2:I5).

Conclusion

Phipps (2007), with a specific focus on grammar teaching and learning, concluded that a part-time Delta he studied did have impact on a teacher's beliefs, awareness, confidence and classroom practice. This detailed study of six teachers' experience of a full-time Delta elaborates on these findings by providing clear evidence – derived from an analysis of coursework, tutor feedback and participant interviews – of a range of ways in which the course impacted on the teachers' knowledge, beliefs, awareness and practices both during and after the course. Overall, the impact of the Delta on these teachers, while variable, was considerable and multi-faceted. A post-Delta observational component to the study would have provided stronger evidence of the links between the course and teachers' subsequent professional practices, but this was not feasible given the resources available for this project (the teachers returned to schools spread out across the UK and, in one case, in Asia). Nonetheless, the insights from this project add to the limited existing literature on the impact which in-service language teacher education can have. In particular, the detailed longitudinal and qualitative evidence of impact presented here is lacking in the literature and this study makes a strong case for the value of such a methodological orientation to the study of teacher learning. The issues highlighted here are of relevance to language teacher educators more generally and will be addressed more specifically in subsequent publications from this project.

Acknowledgements

I am grateful to Cambridge ESOL for funding this project and to the training centre and the participating teachers for their co-operation throughout the study.

References

Borg, M (2005) A case study of the development in pedagogic thinking of a pre-service teacher, *TESL-EJ* 9 (2), 1–30.

Borg, S (forthcoming) The impact of in-service teacher education on language teachers' beliefs, *System* 39 (3).

Bryman, A (2008) *Social research methods* (3rd Edition), Oxford: Oxford University Press.

Busch, D (2010) Pre-service teacher beliefs about language learning: The second language acquisition course as an agent for change, *Language Teaching Research* 14 (3), 318–337.

Denscombe, M (2002) *Ground rules for good research*, Buckingham: Open University Press.

Freeman, D (1993) Renaming experience/reconstructing practice: Developing new understandings of teaching, *Teaching and Teacher Education* 9 (5/6), 485–497.

Goodall, J, Day, C, Lindsay, G, Muijs, D and Harris, A (2005) *Evaluating the impact of continuing professional development (CPD)*, London, UK: Department for Education and Skills.

Guskey, T R (2000) *Evaluating professional development*, Thousand Oaks, CA: Corwin Press.

Hayes, D (Ed.) (1997) *In-service teacher development: International perspectives*, Hemel Hempstead: Prentice Hall.

Kubanyiova, M (2006) Developing a motivational teaching practice in EFL teachers in Slovakia: Challenges of promoting teacher change in EFL contexts, *TESL-EJ* 10 (2), 1–17.

Lamb, M (1995) The consequences of inset, *ELT Journal* 49 (1), 72–80.

Lamie, J M (2004) Presenting a model of change, *Language Teaching Research* 8 (2), 115–142.

Phipps, S (2007) What difference does DELTA make? *Research Notes* 29, 12–16.

Phipps, S (2010) *Language teacher education, beliefs and classroom practices*, Saarbrucken: Lambert Academic Publishing.

Richards, J C, Gallo, P B and Renandya, W A (2001) Exploring teachers' beliefs and the processes of change, *The PAC Journal* 1 (1), 43–64.

Scott, R and Rodgers, B (1995) Changing teachers' conceptions of teaching writing: A collaborative study, *Foreign Language Annals* 28 (2), 234–246.

Zeronis, R (2007), The DELTA revision project – progress update, *Research Notes* 29, 4–7.

Updates on conferences and other events

ALTE events

Some 500 participants from over 50 countries and regions around the world attended the ALTE 4th International Conference in Kraków, Poland in July. Hosted by the Jagiellonian University, the conference provided an opportunity for participants to hear influential voices, discuss key issues and meet colleagues from a variety of different backgrounds. It also gave participants a chance to find out more about the important work ALTE members are doing.

The six plenary presentations headed a conference programme of over 100 papers on key aspects of the theme *'The Impact of Language Frameworks on Assessment, Learning and Teaching viewed from the perspectives of policies, procedures and challenges'*. It was a multilingual conference with presentations delivered in five languages – English, French, German, Polish and Spanish. Cambridge ESOL was well represented at the conference by over a dozen presenters.

One of the highlights of the conference was the LAMI (Language Assessment for Migration and Integration) Forum held under the auspices of the Secretary General of the Council of Europe, Mr Thorbjørn Jagland, and opened by Joseph Sheils, formerly Head of the Language Policy Division of the Council of Europe. The theme of the forum – *Language Testing and Access* – continued the discussions ALTE has been engaged in for a number of years relating to language testing in European migration policy but with a focus this time on the notion of access, in its literal and figurative meanings, and the implications for assessment.

The conference was preceded by a three-day ALTE Extended Learning Course on *The Application of Structural Equation Modelling (SEM) in Language Testing Research*, run by Dr Ardeshir Geranpayeh, Assistant Director, Research and Validation at Cambridge ESOL. The course addressed issues such as exploratory and confirmatory factor analysis, latent variable investigation, multiple regression and building structural models.

Looking ahead to events later in the year, ALTE will run its annual summer testing courses in Copenhagen in September. The first course will be an Introductory Course in Language Testing and will focus on the practical application of testing and assessment theory; and the second course will be an Introductory Course in Research Methodology and will look at research design, research methods, mixed methods, and researching listening and reading.

ALTE's 40th meeting and conference will take place in Bochum, Germany from 16–18 November, and will be hosted by one of ALTE's German members, TestDaF Institute. As at previous meetings, the first two days will include a number of Special Interest Group meetings, and workshops for ALTE members and affiliates, and the third day will be an open conference day for anyone with an interest in language testing. The theme of the conference is *Achieving Context Validity* and the speakers at the conference will include Professor Gillian Wigglesworth (University of Melbourne) and Professor Cyril Weir (University of Bedfordshire). Michael Corrigan from the ALTE Validation Unit will run a workshop on Using Statistical Analysis, and Dr Evelina Galaczi from Research and Validation will give a plenary presentation at the conference entitled 'Investigating the context validity of speaking tests: A case study'.

Together with Lucy Chambers from Research and Validation at Cambridge ESOL, Evelina will also run a two-day Introductory Course on *Assessing Speaking* just after the conference, and Annie Broadhead, Consultant to Cambridge ESOL, will run a one-day Foundation Course on *Language Testing: Getting Started*.

For further information about these events and other ALTE activities, please visit the ALTE website – www.alte.org

To become an Individual Affiliate of ALTE, please download an application form from the ALTE website or contact the Secretariat – info@alte.org

English Profile update

The first quarter of 2011 has seen several English Profile events happening in Cambridge and Switzerland.

Following the publication of the *Common European Framework of Reference for Languages* (CEFR), Cambridge ESOL (in collaboration with other partners such as Cambridge University Press, British Council, English UK) initiated English Profile in 2005, an interdisciplinary research programme which aims to describe the CEFR levels for English. One of the main sources of data for English Profile research is the Cambridge Learner Corpus, a collection of over 40 million words of learner English taken from Cambridge ESOL examination scripts.

On 12 January, around 30 Cambridge ESOL staff attended an internal seminar entitled *English Profile – the CEFR for English.* At this event, one of a regular series of seminars

given by internal and external speakers, Angeliki Salamoura, Nick Saville, Helen Spillett and Stephanie Offord presented an update on the Programme's recent outcomes for Cambridge ESOL staff and explored ways of applying these findings to attendees' own work. The presenters linked the aims of the Programme to the CEFR and presented recent research findings that outline the profile of the English language learner across the six CEFR levels in terms of vocabulary, grammar and functions. Staff attending were able to suggest practical applications of the English Profile findings for teaching and testing, with particular reference to ESOL exams, services and products.

The next event was the 10th English Profile Network Seminar which took place in Cambridge on 10–11 February at the University Centre in Cambridge. This seminar was attended by around 80 delegates from the core group plus the English Profile Network, and it reported the progress of the Programme over the last year in terms of research outcomes and their uses. The first day focused on the year's developments including Dr Tony Green's (University of Bedfordshire) Can Do survey, updates on written and spoken corpora and materials development, a discussion on discourse level aspects of English Profile and a demonstration of the Corpus Query System *Sketch Engine*. We also heard from Professor John Hawkins (University of Cambridge) on new directions for 'Criterial features' and Liz Walter (Cambridge University Press) spoke (on behalf of Annette Capel) about completing the English Vocabulary Profile (formerly known as the English Profile Wordlists). The second day outlined new projects such as the development of a Grammar Profile for English and new methods of analysing English Profile corpus data. Professor Ted Briscoe, Helen Yannakoudakis, Dora Alexopoulou, Oeistein Andersen, members of iLexIR and the University of Cambridge, presented the latest developments in technology and user interfaces and described their aims for future language analysis.

Professor Mike McCarthy (University of Nottingham) rounded up the event (standing in for John Trim, Council of Europe), emphasising the key theme of involving teachers, learners and other professionals in the development of the English Profile Programme. Mike reflected on plans to create a booklet explaining the English Profile Programme and its resources, which will become a practical tool for teachers, materials writers, curriculum designers and other practitioners involved in English Language Teaching.

The next English Profile Network Seminar is due to take place in Brno, Czech Republic in October for education professionals and other interested parties in the region.

On Friday 11 March around 50 delegates attended an English Language Teaching-related 'Cambridge Day' in Geneva co-organised by Cambridge ESOL, Cambridge University Press, IFAGE (Foundation for Adult Education), and ETAS (English Teachers Association of Switzerland). Annette Capel (Cambridge University Press) presented a plenary session on *Constructing the English Profile: A1–C2 Language Descriptors*, providing information on the different strands of the English Profile, focusing on insights gained from the vocabulary research strand that she has carried out over the last four years. Annette presented the English

10th English Profile Network Seminar, 10–11 February 2011, Cambridge

Vocabulary Profile – formerly known as the English Profile Wordlists – under its new name for the first time and showed examples of the new C1 and C2 level data.

To read full seminar reports, find out about upcoming events, or use the new A1–C2 Preview Version of the English Vocabulary Profile, visit the EP website: www.EnglishProfile.org

Conferences

The annual meeting of National Council on Measurement in Education (NCME) took place in New Orleans, from 7–11 April, 2011. Dr Ardeshir Geranpayeh represented Cambridge ESOL in this conference and presented two papers. The first paper, co-authored with Dr Gad Lim, entitled *Standard Setting to an international language framework: findings and challenges*, was presented as part of a co-ordinated session entitled *Standard Setting in an International Context: Issues and Practice*. The authors described a standard setting study and an external validation study to relate the *IELTS* (*International English Language Testing System*), an international exam of English language proficiency, to the CEFR, and offered an explanation for the divergent results obtained by other studies relating various exams to the CEFR. The studies reported by the authors provide insight into and pose questions for the theory and practice of standard setting, including the appropriacy of standard setting where the transitivity of the decision is reversed, the acceptability of divergent standard setting outcomes in such contexts, and the application of external validation to verify the results of standard setting.

Ardeshir also presented another paper, co-authored with Dr Muhammad Naveed Khalid, entitled *Detection of Differential Item Functioning and Scale Purification*. The authors argue that item bias or differential item functioning (DIF) has an important impact on the fairness of psychological and educational testing. In this paper, DIF is seen as a lack of fit to an item response theory (IRT) model. They argue that inferences about the presence and importance of DIF require a process of so-called test purification where items with DIF are identified using statistical tests and DIF is modelled using group-specific item parameters. A stepwise procedure is proposed where DIF items are identified one or two at a time. Simulation studies are presented to illustrate the power and Type I error rate of the procedure. The authors also proposed a method for defining a stopping rule for the searching procedure for DIF. The estimate of the difference between the means and variances of the ability distributions of the studied groups of respondents is used as an effect size and the purification procedure is stopped when the change in this effect size becomes negligible.

Participants in the Standard Setting in an International Context session (left to right): Susan Davis-Becker, Chad Buckendahl, Michael Rodriguez, Mary Pitoniak, Ardeshir Geranpayeh, Greg Cizek and Leslie Shaw