CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

# Research Notes

## Issue 57

## August 2014

ISSN 1756-509X

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

# Research Notes

Issue 57 / August 2014
A quarterly publication reporting on learning, teaching and assessment

*Production Team*
Karolina Herman, *Marketing Project Co-ordinator*, Cambridge English Language Assessment
John Savage, *Publications Assistant*, Cambridge English Language Assessment

Printed in the United Kingdom by Canon Business Services

# Research Notes

Issue 57

August 2014

## Contents

# Editorial

Welcome to issue 57 of *Research Notes*, our quarterly publication reporting on matters relating to learning, teaching and assessment within Cambridge English Language Assessment. This issue mainly presents research projects undertaken within the 2013 round of the Cambridge English Funded Research Programme. Through this programme we support world-class research into many theoretical and practical aspects of our products and services, including aspects of test design, delivery and their impact around the world.

In the first article Ruth Breeze and Hanne Roothooft report on their investigation into school teachers' views on the impact on classroom practice of *Cambridge English: Young Learners* exams in Spanish primary schools. They carried out a qualitative investigation of the effect of using *Cambridge English: Young Learners* in Navarra through interviewing teachers about their reasons for incorporating these tests in the curriculum, their attitudes to this development, and the ways in which preparing students affected their classroom practice. This article provides a generally positive view of the integration of these exams into primary education in Spain whilst noting some resistance which other studies should explore.

Next, Liyan Huang and Angelo Papakosmas explore the impact of a teaching qualification – the *Teaching Knowledge Test* (*TKT*) – on Chinese teachers' beliefs, knowledge and practice using focus groups and questionnaires. They explored the scope and intensity of the impact of taking Modules 1, 2 and 3 of *TKT* on over 200 Chinese teachers from a range of educational sectors, which enabled them to suggest the role that various contextual and demographic factors have in shaping *TKT* impact in China and the broadly positive impact of *TKT* on teachers, including their own language development.

Although the focus of this issue is the 2013 round of the Cambridge English Funded Research Programme, a 2012 study of the impact of the *Cambridge English: First* test in Cyprus provides a useful complement to the other findings presented here. In this article Dina Tsagari explores the face validity of *Cambridge English: First* based on teachers' insights into the effects of this test on their teaching and their students in private language schools in Cyprus. This study aimed to understand teachers' perceptions of the utility, efficiency, practicality, difficulty, content validity and washback of *Cambridge English: First* through in-depth interviews with teachers. The author reports several unintended impacts of this exam and reiterates that seeking feedback from stakeholders is a crucial part of the assessment process.

The subsequent two articles return to 2013 funded research. First, Ana María Rozzi, Verónica Pinto, Marina González and Yanina Crimi report on the impact of a range of Cambridge English exams on institutional change in four language schools in Argentina, an aspect of impact that is under-researched. They investigated why the examinations were adopted and whether and how this had informed pedagogic, administrative and managerial areas. Head teachers and class teachers were interviewed and completed questionnaires, whilst premises and school documents were inspected using checklists. An informative picture emerges from this research, which shows how institutional change is a complex and slow-moving phenomenon in any context.

This is followed by an article by Okim Kang and Linxiao Wang, who look in detail at candidate speaking performances at different CEFR levels, reporting on the impact of task types and identifying features of interaction that distinguish lower from higher levels. They report on their corpus-informed study that compared the linguistic features of candidates' output in an individual speaking task and a paired task in *Cambridge English: Preliminary, Cambridge English: First, Cambridge English: Advanced* and *Cambridge English: Proficiency* exams. They found distinctive differences in linguistic features in the two tasks across the four levels and between individual and paired tasks, which has implications for Speaking test design, rating scale development and examiner training, as well as implications for the classroom.

Staying with speaking but considering examiners rather than candidates, the final article in this issue, by Sue Gilbert and Georgia Staub, presents an investigation into experienced Speaking Examiners' perceptions of the accuracy of their assessments of *Cambridge English: Advanced* Speaking tests. This study was commissioned by Cambridge English to gain insights into how examiners in Switzerland perceive the reliability of their own assessments and the factors which may impact on this. The article provides a unique insight into established examiners' perceptions and will feed into future Speaking test design and examiner training and professional development.

Our Funded Research Programme continues to provide important insights into examiner and candidate behaviour, stakeholder perceptions and more theoretical studies into test design and rating. We hope that this issue, along with issues 47, 52 and 54, encourages others to consider submitting a proposal for funded research in the future. The Call for Proposals is publicised every August online at www.cambridgeenglish.org/research-and-validation and can be found at the end of the issue.

# Teacher perspectives on implementing Cambridge English: Young Learners exams in Spanish schools

**RUTH BREEZE** UNIVERSITY OF NAVARRA, SPAIN
**HANNE ROOTHOOFT** UNIVERSITY OF NAVARRA, SPAIN

## Introduction

In the last 15 years, the teaching of English has been established as a priority in Spanish schools. Although national legislation only requires a foreign language to be taught from the age of 8 onwards, most of Spain's autonomous communities, which have the authority to regulate non-basic aspects of education within their own areas (Spanish Eurydice Unit 2009), have established programmes by which all children begin learning English at the age of 6, or even 3. For example, the autonomous community of Madrid has set the goal for the majority of secondary school students to reach a B2 level in English before leaving school (Ashton, Salamoura and Diaz 2012). Private and subsidised schools have long been eager to show that their students achieve a high level of English, and state schools are now joining in, with large networks of state schools participating in state-sponsored programmes that offer an increased emphasis on English and content taught through English in the primary curriculum. Against this background, it is hardly surprising that an increasing number of candidates are taking *Cambridge English: Young Learners* exams in Spain.

Some recent impact studies, such as that by Ashton et al (2012), offer a comprehensive overview of the effects which *Cambridge English: Young Learners* is having on bilingual schools in particular areas of Spain, looking at what the school management intended, and how the implementation of the exams has been perceived by teachers, learners and parents. However, the published findings of this study have certain limitations. First, the results of this research are mainly obtained through questionnaire data, which entails the risk that teachers' own opinions are not accurately reflected. Secondly, the study focuses on the Bilingual English Development and Assessment (BEDA) project carried out by Cambridge English in conjunction with the Federation of Religious Schools in the Madrid area (FERE Madrid) since 2008. It therefore covers one specific programme in which Cambridge English was involved from the outset (Blanco and Nicholson 2010), and is not entirely representative of the country as a whole.

Against this background, the aim of the present study was to carry out a qualitative investigation of the effect of using *Cambridge English: Young Learners* in Spanish schools outside Madrid by focusing on a set of key players within the school system, namely the teachers. Our research was designed to focus specifically on one set of stakeholders, the teachers, and most particularly on the way putting candidates forward for the *Cambridge English: Young Learners* exams has influenced their classroom practice. For this reason, throughout this article we will use the term 'washback' to describe the effect of the exams on teaching and learning as reported by our subjects (Cheng 2005). Although classroom washback is undeniably part of the impact which these exams have, the notion of impact, as defined by Saville (2012), includes many other aspects which could also be understood to form part of the ongoing effects of the tests within a particular social context. We felt it was particularly useful to understand the teachers' perceptions and attitudes concerning the *Cambridge English: Young Learners* exams. Teachers often play an active part in deciding whether or not a particular school is going to prepare candidates for the exams, and always have a major role in determining the type of preparation that students are given. On the other hand, the teacher is also often the person who feels most threatened by the notion of introducing external assessment of their students' level: if students fail to perform well on the test, this may reflect badly on the teacher. The teacher is thus a key player in the move towards incorporating external exams in the school curriculum, and it is important to understand their perspectives on this process, and to comprehend their needs. We chose the community of Navarra as it offered a suitable setting for a study of this kind, since *Cambridge English: Young Learners* candidate numbers have grown steadily over the last 15 years, in parallel to the gradual expansion of English across the school system.

In view of this, we designed a qualitative study to investigate teacher perceptions of the reasons for incorporating *Cambridge English: Young Learners* in the primary English curriculum in Spain, their attitudes to this development, and the ways in which preparing students for *Cambridge English: Young Learners* affected their classroom practice. This study therefore focused on teacher perceptions and attitudes, and most particularly on teachers' accounts of washback effects.

## Research design and method

This section sets out the final design of the project as a whole. In this section we include a brief summary of the research design and plan, and explain some adaptations that were made. We then describe the process of conducting the interviews and analysing the data. Our principal aims in designing this project were to establish:

- why teachers think schools choose to incorporate *Cambridge English: Young Learners*
- what their attitudes are towards this decision
- what effect the incorporation of *Cambridge English: Young Learners* has in the short and long term on their classroom practice.

We therefore opted to use qualitative methodology based on thematic analysis of interview data obtained in the local context. Our completed study is based on the analysis of a series of semi-structured interviews with 22 primary school teachers working in different types of schools in Navarra, Spain. We adapted an interview template developed previously by Cambridge English Language Assessment to draft a schedule that would be suitable for use in the context of *Cambridge English: Young Learners* in local primary schools, and prepared Spanish and English versions. The final interview schedule (see Appendix 1) is organised into three parts, reflecting the three research questions: teacher perceptions of the decision to adopt *Cambridge English: Young Learners* and their attitude towards this decision (parts 1 and 2); and the washback effects of using *Cambridge English: Young Learners* in the classroom (part 3). The initial version of the whole interview schedule was trialled with three primary school teachers, but since it did not appear to pose any difficulties, and was based on the Cambridge English template that had already been used elsewhere, we decided that extensive piloting was not necessary. Once the interview schedule had been approved, we began to apply this schedule with a sample of primary school teachers working in 13 local schools (22 subjects) (see Appendix 2, Table 1). A total of 22 teachers were interviewed by the two researchers, in sessions which lasted 20–40 minutes. All of the interviews were digitally recorded and then transcribed. In the second phase of the research, both researchers read the transcriptions independently in order to determine the main first-order constructs that emerged in response to the three research questions. After meeting to discuss these constructs and resolve discrepancies, they then re-read the transcriptions to identify sub-dimensions and examine how these were articulated with the main constructs.

### Schools and teachers

The 'Comunidad Foral de Navarra' is an autonomous community in northern Spain situated between the Basque Country, Aragon and La Rioja. It has around 600,000 inhabitants, over half of whom live in or around the capital, Pamplona. The region is generally considered to be one of the wealthier areas of Spain, and the quality of education in the area is high. Apart from one private school which was established in 2011, all the schools in the area are subsidised by the state. At primary level, these schools fall into two main categories, namely state schools (*colegios publicos*), and church schools or collectives (*colegios concertados*). The former are entirely funded by the state, whereas the latter receive state funding that covers most of the teachers' salaries, but rely on parents' contributions to pay other costs. The teachers in this study worked in 10 schools within the *concertado* system and three state schools. This sample is thus representative of the schools which provide most of the candidates for Cambridge English exams in the area covered by Exam Centre ES007, which is the only open centre operating in the Comunidad Foral de Navarra. The emphasis on English within the curriculum varied greatly from one school to another, with some schools only devoting 4 hours a week to English language classes, while others had around 50% of the curriculum in English, including 4 hours of English as such, plus up to 8 hours of other subjects (arts and crafts, science, music, social sciences, or physical education) taught in English (see Appendix 2, Table 2).

## Results

The researchers identified the following areas of interest:

### Reasons why schools adopt Cambridge English: Young Learners

The teachers interviewed were all asked why, in their opinion, the decision had been made to adopt *Cambridge English: Young Learners* exams in their school. In this section, our analysis of the results will focus first on the extensive accounts by teachers from the *concertado* sector, before providing a brief description of the picture in the three state schools where the decision to offer *Cambridge English: Young Learners* had proved much more controversial.

First, it has to be pointed out that the teachers often stated that they did not know why the school had decided to adopt *Cambridge English: Young Learners*, or that they themselves had not been party to the decision, because it dated from before their time, or because it had been made by the school's management without consulting them. Despite this, almost all the teachers volunteered their own opinion as to why their school had decided to offer *Cambridge English: Young Learners*.

Many of the teachers interviewed felt that the school had decided to offer *Cambridge English: Young Learners* in order to present a positive image to the general public and, particularly, to current and prospective parents. In most cases, the schools were also involved in projects to improve their English teaching at primary level, and *Cambridge English: Young Learners* seemed to fit well with this objective. The teachers explicitly linked the decision to adopt *Cambridge English: Young Learners* to the school management's desire to show stakeholders that the school had a high level of English, and in many cases to substantiate the claim that the school was 'bilingual' or 'international'. Competition with other schools within the same sector appeared to be an important factor. The choice of Cambridge English exams rather than other available exams was explained in terms of the 'prestige' of Cambridge English exams, or the importance of 'international' certification.

Teachers also frequently stated that schools were looking for external validation of an objective nature, and that Cambridge exams met that need. Teachers expressed the idea that management/parents/other stakeholders appreciate the need for impartial validation of the school's English teaching: by obtaining good results in *Cambridge English: Young Learners*, the school would be able to measure its pupils' level of English objectively, and thus also evaluate how well their English language teaching was functioning. They also made two other important points. First, the fact that *Cambridge English: Young Learners* is linked to the Common European Framework of Reference for Languages (CEFR) and therefore uses a recognisable yardstick means that students are tested in a way that makes sense beyond the immediate context. So if students reach *Cambridge English: Flyers* (A2) by the end of primary education, it is reasonable to suppose that secondary schools can start the curriculum for the B1 level. Secondly, the fact that *Cambridge English: Young Learners* tests measure pupils' levels in all four skills makes them more appropriate than the local English tests (Gobierno de Navarra 2013).

Teachers mentioned the need to prepare children for future exams of a similar kind, that is, not just exams given by the school itself, but external competitive exams. It was clear talking to many of the teachers that they tended to see *Cambridge English: Young Learners* as the first rung on the ladder, as something that would help to socialise children into a testing system: as one teacher commented, 'it's the first step, but it doesn't stop there'.

**Mixed attitudes in state schools**

It should be noted that the teachers from the three state schools told a very different story in answer to this first question. In all three state schools, the initiative to prepare children for *Cambridge English: Young Learners* and offer the exams in the school had been taken by the teachers themselves, although in all three cases these teachers mentioned that the school management supported the idea. In all cases, these were teachers who had been used to working with *Cambridge English: Young Learners* in other settings and who had found the exams useful and attractive. Two of these teachers reported major difficulties with colleagues, since there was considerable resistance to the idea of introducing an exam which: a) was not run by the Spanish authorities; b) had to be paid for by parents; and c) therefore might create divisions within the group of children (see further discussion of this point below). In one of the state schools, the head was interested in implementing *Cambridge English: Young Learners*, but was encountering various difficulties, including the lack of permanent staff, the expense for parents, and the idea that the teachers' freedom would be constrained by having to offer preparation for the exams. In another, resistance came from people responsible for teaching extracurricular English, who perhaps felt threatened by the mainstream English teacher taking control of preparing students for *Cambridge English: Young Learners*. In the third case, opposition came from colleagues within the English department who resisted the notion of putting students in for what they termed a 'private' exam. From talking to the teachers who were interested in their students taking the exams, it became clear that teachers in the state system regard the idea of an external exam as both a threat and an imposition. In particular, the notion of a 'commercial' exam, that is, an exam perceived (however erroneously) to be part of a money-making concern, was thought to be distasteful. On the other hand, head teachers appeared to be interested in the possibility, both as a source of external validation for the school's English programmes and as evidence to offer parents.

**Teacher attitudes towards Cambridge English: Young Learners**

All of the teachers expressed opinions and attitudes concerning the exams themselves (see Appendix 3, Tables 3 and 4). These fell into five main categories:

- comments on useful or attractive attributes of the exams themselves
- comments on practicalities concerning ease of preparation, availability of supplementary material, etc.
- usefulness of the exams as an evaluative instrument to measure attainment and diagnose problems, in terms of both learning and teaching
- importance of the exams for motivating students
- the role of the exams in motivating teachers.

*Attractive features of Cambridge English: Young Learners*

Most teachers particularly valued the fact that *Cambridge English: Young Learners* gives equal weighting to the four skills, because they felt that this was the most appropriate framework for English teaching in general. They contrasted *Cambridge English: Young Learners* favourably with the exams set by the regional authorities for all children in the fourth year of primary school (age 9–10) which test only reading, writing and listening. However, it should be noted that a few teachers expressed a contrasting viewpoint, namely that primary English should focus on speaking and naturalistic interaction, and that it was not useful to spend so much time on reading and writing. On the whole, most teachers particularly valued the emphasis that *Cambridge English: Young Learners* put on oral skills, and emphasised that it was useful that children knew these skills would be tested, and that they would be examined by an external examiner. They felt that the fact that students can have a conversation in English with a person whom they do not know helps to build up their confidence. One teacher also emphasised the good fit between *Cambridge English: Young Learners* tests and what teachers ought to be doing in the classroom.

*Practical to prepare for and easy for children to do*

A large number of teachers mentioned that they liked the format of the exams and found them easy to prepare children for. One teacher specifically mentioned the 'child-friendly topics', and several stated that the activities were designed in a way that fitted with the ability and interests of the age group in question. One teacher commented that the exams were 'so well thought-out'. The teachers clearly valued the transparency of *Cambridge English: Young Learners* (handbooks which set out the criteria clearly, availability of past papers on the Cambridge English website, availability of other preparation material and ideas), and particularly emphasised the usefulness of the *Cambridge English: Young Learners* wordlists as giving a clear goal as to what students would be expected to know.

*Measure achievement and diagnose strengths and weaknesses of individuals and groups*

Students' achievements on *Cambridge English: Young Learners* (the exams themselves or practice tests) are a good indication of their progress, measured against an objective external standard. One teacher said that *Cambridge English: Young Learners* set a standard in that they 'show us what students should know'. Many teachers considered that when students get four or five shields in *Cambridge English: Movers* or *Cambridge English: Flyers*, this shows that they have reached the corresponding level on the CEFR, and this helped them to position students and groups with regard to past and future learning. Since many schools offer *Cambridge English: Flyers* at the end of primary school (age 11–12), some teachers commented that this provides evidence that they should start secondary school working towards the B1 level. This is an important point for gauging the appropriacy of programmes in primary school, as well as for evaluating the performance of groups or individual students and comparing achievement from one year to another. In this sense, exam results can also be used for teachers' self-assessment, or for assessment of the English curriculum: as one teacher said, 'the exams help us to evaluate ourselves and the way we teach'. On the other hand, it was also important for some teachers that the results can be used to diagnose individual students' strengths and weaknesses.

*Motivation for students*

According to the teachers in this sample, students are usually highly motivated to do *Cambridge English: Young Learners* exams, and this has a good effect on their attitude and behaviour. Students may also occasionally become too competitive or too nervous, in which case the effect may be less positive, but as will be discussed in more detail below, most teachers do not see this as a problem. In general the idea of external assessment makes the students work harder and pay more attention in the classroom. Slower students are also motivated by the idea that they will get something from doing the exam. In fact, since schools often prefer students to take the exam when the vast majority of pupils are ready to do it well; this tends to promote equality among the students, and even the weakest students feel satisfied by their results. Parents are generally very keen for their children to take the exam, which adds to the motivation factor. In some schools, however, not all the students take the exam and teachers often make recommendations to parents about whether or not their children should take it.

*Motivation for teachers*

Many teachers also feel motivated by the idea that their pupils are going to take an external exam. Although at first they may perceive this as a threat, in the long term they generally express positive opinions about the idea that their students are going to take external exams such as *Cambridge English: Young Learners*. This apprehension was only voiced by teachers in schools where the exams were just being implemented into the system. In schools with a longer track record with *Cambridge English: Young Learners*, this did not seem to be a problem. On the whole, despite what might be perceived as extra work, teachers expressed a sense of satisfaction in their pupils' achievements. Moreover, they themselves often expressed a sense of pride and pleasure in being part of what they termed the 'Cambridge team': they understood the 'Cambridge' label to be highly prestigious, and felt that the fact that their students did well in Cambridge English exams was a confirmation of their own expertise as teachers.

**Washback**

Teachers answered a large number of questions designed to elicit ways in which putting students in for *Cambridge English: Young Learners* might affect classroom practice. Interestingly, their initial response was often denial that *Cambridge English: Young Learners* exams had any influence on their teaching. However, when prompted further, most of the teachers came up with changes of emphasis that had come about as a result of the exams, or with concrete examples of activities that they would not have done if they had not been preparing students for *Cambridge English: Young Learners*. The washback effects are set out schematically in Appendix 3, Table 5.

*No specific washback effect*

Several of the teachers interviewed expressed opinions along the lines of 'it doesn't affect what we do in the classroom', or 'I would teach that way anyway'. Some teachers stated that 'it fits very well with what we are doing anyway', 'the books we use have a similar approach', or 'there is nothing in the exam preparation we wouldn't do in class anyway'. Rather than invalidating the notion that *Cambridge English: Young Learners* exams have a significant washback effect, this type of statement provides evidence that *Cambridge English: Young Learners* exams probably reflect many aspects of state-of-the-art practice in primary school English. In fact, it was observable that schools with very successful primary English programmes perceived less washback than those where the teachers were still struggling to get the English component into shape. This is an important point, because it appears to illustrate the excellent fit between *Cambridge English: Young Learners* and current thinking and good practice in primary school English teaching.

*Change coursebook*

Putting students in for *Cambridge English: Young Learners* exams had made some teachers see that their existing coursebook was not adequate, and had prompted a change to more up-to-date material.

*Move towards a greater balance between skills*

Implementing *Cambridge English: Young Learners* had led some teachers to change the balance between skills, although the actual effect of this varied according to what the previous situation had been. In some cases, it meant that teachers had started to work more on oral skills, and in this area, some had now started to use pairwork in the classroom. One teacher said that she continued to teach speaking as she had done before, but that she now paid less attention to the mistakes that students made, because she knew that Cambridge did not penalise students for these as long as they managed to communicate successfully. In other cases, the adoption of *Cambridge English: Young Learners* meant that teachers introduced reading and writing (particularly spelling) before they would otherwise have done so. Whether the effect was that teachers taught more speaking, or more reading and writing, the overall net impact of *Cambridge English: Young Learners* in many cases was that schools moved over to a greater balance between skills. As one teacher said, 'the balance of skills is better now'.

*Specific additions or changes in emphasis*

On many occasions, some particular features related to *Cambridge English: Young Learners* exams were singled out for attention. One teacher mentioned the positive effect she had experienced when she started using the Cambridge listening recordings as a source of different voices and accents. Another said that she now particularly stressed prepositions of place, because she knew that the exams generally required students to understand and describe where things were in relation to each other. Another teacher mentioned emphasising the past simple tense, since he knew that students would need it for the exam. There was considerable variation in the responses to this question, but it was noticeable that a large proportion of the teachers interviewed specifically mentioned the *Cambridge English: Starters, Cambridge English: Movers* and *Cambridge English: Flyers* wordlists as providing the essential words they would use in classroom activities, and which they would ensure that their pupils learned.

*Addition of specific exam practice sessions*

In most schools, teachers reported using one hour-long session per week for the whole year, or using the second and third term to do past papers and specifically prepare students for *Cambridge English: Young Learners* exams. They said that this was a positive experience, because it provided them with 'new tools', or 'a different way of teaching'. They reported no problems with this because of the abundance of material and support available, in the form of past papers, resource books and web material. In a few of the schools, a teaching assistant or student teacher was given the task of preparing the students for the Speaking test: one teacher mentioned that *Cambridge English: Young Learners* material actually provided 'a more structured approach for these sessions'.

*Influence on the way teachers assess students*

In addition to actually putting candidates in for *Cambridge English: Starters, Cambridge English: Movers* and *Cambridge English: Flyers* and studying their results, teachers also use *Cambridge English: Young Learners* past papers to evaluate their students, and are influenced by the *Cambridge English: Young Learners* model and format when writing their own tests. Even teachers who do not directly prepare students for an exam in the levels they teach state that they use parts of the Cambridge English exams, for instance the Listening paper, in their classes because they think the material is good. One experienced teacher commented that 'because they're so well thought out, you often think that when you're making an exam for General English you think of the way that they do it in *Cambridge English: Flyers* and you use, maybe adapt, those ideas'. There is thus evidence that the *Cambridge English: Young Learners* model provides a robust framework for assessment in primary schools which teachers value highly, because it is sound in its evaluation of students' language skills, yet leaves some scope for the teacher's own creativity.

*Teachers give more feedback*

A few teachers mentioned providing better, more detailed feedback to students on the basis of *Cambridge English: Young Learners* past papers or other exercises in the style of *Cambridge English: Young Learners* exams. One teacher alluded to the surprises that she sometimes had, when she found that a seemingly weak student had good results on Listening tests, or passed the Speaking test: 'it makes you see your students a different way'. This was particularly important in that it provided insights into achievements in areas other than reading and writing skills, which may tend to be easier to test in the classroom. Another teacher mentioned that using *Cambridge English: Young Learners* material and handbooks had enabled her to 'give students feedback on speaking', which she would not have been able to do before since she had not known what criteria to use.

*More work in one sense, but less in another*

Some teachers stated that having to make copies of past papers and correct them could be understood as giving them a greater workload. However, this was easily outweighed by the fact that there were many freely available exam-related activities that they could use (e.g. on the website). They also noted that *Cambridge English: Young Learners* exams are easy to correct, since most of the answers are multiple choice or single-word options.

**Challenges of preparing students for Cambridge English: Young Learners**

As we explain above, all the teachers involved are positive about the format of the *Cambridge English: Young Learners* exams and can see many more advantages than disadvantages in preparing their pupils for the exams. Nevertheless, they also reported a few drawbacks emerging in the context of offering the exams within a school (see Appendix 3, Table 4), which are analysed below. Difficulties with the actual exam format are described later on.

*Economic issues*

The most important disadvantage is the money that parents need to pay. Some teachers mentioned that occasionally parents do not want their children to take the exam because they find them too expensive. In certain schools where the children were from more modest backgrounds this was a particularly difficult point. Especially for the state schools, the fact that the exams are not free can be a problem. One teacher working in a state school said that her colleagues are opposed to the exams because they think they are commercial.

*Nervous students*

Another possible problem concerns the students. Even though all the teachers agree that the *Cambridge English: Young Learners* exams are very motivating for the children, several teachers also notice their students can get quite nervous. However, this is not necessarily seen as a problem by most teachers. One teacher thinks it's 'a preparation for life, you always get nervous before an exam. It's natural.'

*Weak students*

Another concern related to the students has to do with the weaker students who might not do very well in the exam. A few teachers seemed worried that some students might get disappointed when they get the results: 'I think a disadvantage will be the kind of students that will never be able to take them [the exams], because they are going to get frustrated'. In most of the schools in this study, the exams are optional, which often results in only the strong students taking the exam. Several teachers tend to discourage weaker students from enrolling on the exam.

*Students are too young*

Some teachers also thought that primary children are too young to take standardised exams. This was especially a problem in one school, where children are taking the exams at an earlier age than in other schools. However, this is a special case and the age factor was not considered to be a problem in other schools. As discussed earlier, one teacher even stressed the child-friendliness of the exams.

*Low levels*

The low level of *Cambridge English: Young Learners* was mentioned by several teachers. It is interesting that many schools offer *Cambridge English: Flyers*, but fewer schools offer *Cambridge English: Movers* and even fewer have *Cambridge English: Starters*. One teacher who saw the *Cambridge English: Flyers* exams as something very positive stated that it would not be such a good idea to do *Cambridge English: Starters* and *Cambridge English: Movers*, as this might put more pressure on teachers and make weaker students feel bad about their English. Another teacher did not see many advantages of doing *Cambridge English: Movers*, if two years after that the students would be taking *Cambridge English: Flyers*. The fact that the level of these exams is quite low compared to exams such as *Cambridge English: Preliminary* and *Cambridge English: First* was mentioned by a few teachers and it also appears some parents do not wish their children to take exams in primary school because they prefer to wait until their children can take higher-level exams. It was noticeable that in this, our sample appeared to run counter to the patterns found in other areas of Spain, where numbers are roughly similar between the three levels.

*Difficulty of preparing for the Speaking test*

Because the classes usually have 25 to 30 pupils, some teachers mention the specific difficulty of preparing students for the Speaking test. In certain schools, this problem was solved with the help of teaching assistants, but these are not available in all of the schools: 'Preparing an individual mock oral exam takes a lot of time when you've got 27 students in the classroom. That is what I think is most difficult.'

*Extra work and pressure on teachers*

Some disadvantages for the teachers themselves include some extra work, but as mentioned earlier this is not a real problem for most teachers. However, when discussing this, several teachers also added the point that it is usually not difficult to prepare for these exams, because there is abundant material available, and that it is easy to correct past papers. A further disadvantage for the teachers was the issue of pressure. Some teachers also admitted to feeling a certain pressure to produce good results: 'Nobody tells you they have to do well in order for the school to have a good reputation or whatever. But you always know that it's something important.'

*Possible solutions*

During the course of this study, the interviewers found that establishing stronger channels of communication between the local Cambridge English centre and the teachers who were giving preparation classes was in itself a positive step towards solving some of the issues that arose. Through the interviews, some teachers became more aware of the abundance of material available on the various Cambridge English websites. Others expressed an interest in what other teachers had said, and further contacts were established within the network of schools preparing students for *Cambridge English: Young Learners* (for example, one school wanted to know how other schools used the teaching assistants, while in another case, a school asked for a teacher with experience to give a seminar to show its staff how to prepare students for the Speaking test). In general, professional networks of this kind on a local level, which arise informally through training seminars and conferences, and may then be formalised through newsletters and special events, appear to offer support and guidance for teachers or schools that can help them to persevere when problems arise.

**Difficulties with specific parts of Cambridge English: Young Learners**

The teachers in this sample identified a number of specific aspects of the exams, discussed below.

*Speaking*

The story in the Speaking test for *Cambridge English: Movers* and *Cambridge English: Flyers* is perceived to be particularly difficult for children to do, and challenging to prepare for. Although teachers agreed that primary school pupils relate easily to narrative, because it is the genre to which they are most accustomed, several of the professionals in this sample singled out the story task as being the most difficult, and indicated that they would like more support and material for preparing pupils for this task.

The purpose of the question and answer activity in *Cambridge English: Flyers* was not well understood by one of the teachers, who felt that it was not appropriate to her students' level of achievement.

*Reading, Writing and Listening*

Two of the interviewees identified the Reading and Writing paper as being the most difficult part of the exam for their students. In one case, at a school in which the students take *Cambridge English: Movers* in the second year of primary, that is, age 6–7, the teacher reported that the information transfer exercises in the Reading and Writing paper (Part 5) posed special difficulties for the students. However, this would seem to be explicable in terms of the students' developmental level, and the fact that some of them still have difficulties with basic literacy skills in their own language at this age. In another case, a teacher commented on the difficulty of the last three sections in the *Cambridge English: Flyers* Reading and Writing paper, and mentioned the need for more guidance as to how to prepare students for these parts of the exam. One teacher also commented on the attention span required for the Listening test in *Cambridge English: Movers* and *Cambridge English: Flyers*, which seemed rather long for her students.

## Discussion

It is clear that the teachers involved in this study all have a very positive attitude towards *Cambridge English: Young Learners* exams. Any apprehension that they might have felt when starting out with these exams was usually dispelled after they had discovered the wealth of material available and experienced the positive effects on student motivation. Of course, since the teachers in our sample are, by definition, teachers who have agreed to prepare students for the exams, our sample may not be representative of the teaching profession in the area as a whole, and a survey of all teachers involved in primary school English would probably produce rather different results. Nonetheless, it is encouraging that these teachers reported in several cases that they or their colleagues had changed their minds once they learned more about the exams and the available resources.

Regarding the specific points identified in our analysis, it must be said that they are very much in line with the results of other published studies. As for reasons why the schools had adopted *Cambridge English: Young Learners*, the points made by our respondents coincided largely with the findings obtained by Ashton et al (2012) in their preliminary assessment of the impact of the BEDA project in Madrid. These authors identified three main factors: keeping up with other schools, making the school's achievements in English more visible to parents, and having an external reference against which they could measure the students' improvement. In our case, we put the first two of these factors together as 'improving the school's image', a broad notion encompassing both competition with other schools and showing that the school is 'bilingual' or has a high level of achievement in English, mainly for the purpose of persuading parents to send their children there. The point about external validation is also regarded as extremely important by the teachers in our sample, who see the exams as a useful yardstick that makes it possible to determine course objectives, measure children's achievement, and check that their own programmes are working effectively. Conversely, resistance to these exams in the state sector was often explained by the idea that Spanish state schools should not have external assessment, or only that provided by other entities within the Spanish state sector. Finally, on the point as to why schools chose to adopt *Cambridge English: Young Learners*, many of the teachers felt that this was done at primary level with a long-term strategic objective in mind: this was the first step on a ladder which would lead to students taking *Cambridge English: First* or other international exams much later on. For *Cambridge English: Young Learners* exams to become more widely accepted in this sector, it is necessary to show what these exams offer teachers, in terms of resources, student motivation, and so on. It might perhaps also be useful to explain that some examination providers are non-profit-making entities that are run for educational purposes, since this concept is not familiar to them.

Concerning teachers' attitudes, our study reflects many of the results obtained elsewhere, as well as adding some new points to the discussion. The teachers in our sample reported an improvement in student motivation, which is consistent with the findings of Ashton et al (2012), as well as with studies of the impact of *Cambridge English: Young Learners* from other areas in the world (Chambers, Elliott and Jianguo 2012, Khalifa, Nguyen and Walker 2012, Salamoura, Hamilton and Octor 2012). Most of the teachers in our study also reported an improvement in their own motivation, which paralleled the findings by Ashton et al (2012), and in other studies from both the private and the public sector in very different settings (Chambers et al 2012, Khalifa et al 2012). It therefore appears that the extrinsic motivation added by the fact of having to prepare for an external exam is regarded as positive, and does not in any sense have a negative impact on the teachers' professional satisfaction. The teachers in our study also greatly appreciated the balanced design of the exam, which gives importance to all four skills, and its attractive, child-friendly design. As Ashton et al (2012) found, our teachers reported that children related well to the topics and tasks. Our teachers also used the exams for diagnostic purposes, to contrast different groups and years, to identify individual differences, to give feedback, and to gauge where support should be given. Other points emphasised by our sample of teachers were the practicality and usefulness of the exam design itself, and of the material available to help prepare for the tests. The abundance of resources (past papers, resource books, web-based material) was obviously a major factor in reinforcing the teachers' positive attitudes towards the exams.

As far as washback on classroom practice was concerned, it must be said that almost all the comments made by the teachers seemed to indicate that using *Cambridge English: Young Learners* had had a positive impact on their teaching. Those who began by stating that the exams had no effect at all generally explained this by saying that the exams fitted perfectly into their existing practice. If these teachers did not need to modify their practice too much, this tends to suggest that the exams are a reflection of current good practice in primary English. On the other hand, many teachers did report changes in their practice as a result of using *Cambridge English: Young Learners*. Some reported that this had led them to change their coursebook to a more up-to-date one which incorporated exam practice material. Many teachers reported that using the exams had brought about a change in emphasis, either by giving more weighting in the classroom to one or two of the four skills which had previously been neglected, or by adding some specific sub-skills that they had not taught before. Although the innovations varied from school to school, in general the changes that occurred as a result of preparing students for *Cambridge English: Young Learners* would seem to point towards a move towards a more balanced curriculum: teachers who had formerly only focused on speaking and listening in the context of class projects began to include reading and writing as well, while schools where the main focus had been on written skills started to do more speaking practice in the classroom. Although these findings contrast to some extent with those of Ashton et al (2012) and Chambers et al (2012), who conclude that the introduction of *Cambridge English: Young Learners* meant that the focus on oral skills in the classroom intensified, this discrepancy is probably a reflection of differences in previous practice: in our area, some schools had been working on class projects that involved considerable use of spoken language before they took up *Cambridge English: Young Learners*. For them, the Reading and Writing parts of the exam required a change in focus. For other schools, where grammar and vocabulary or reading and writing had determined the core course content, the need to practise speaking had been perceived as an important innovation. On the other hand, the need to practise the actual exam format inevitably leads to spending a certain amount of time on test-completion techniques (see Gu, Khalifa, Yan and Tian 2012:46). However, on the whole we might conclude that the overall influence of *Cambridge English: Young Learners* lies in ensuring a balanced curriculum in primary English, in which time and effort are dedicated to each skill. The effect of these exams, and of other important influences such as use of the

CEFR descriptors to define the objectives for primary school English, is to bring about greater homogeneity and balance across the programmes offered by different schools.

Three other points worthy of note emerged from the teachers' reports on changes in classroom practice. First, the wordlists were mentioned by almost all the teachers as providing the essential backbone of the lexis taught in primary English. This again tended to bring schools closer to each other in terms of what is taught. Second, the need to practise the format of the Speaking test enabled teachers to make more productive use of teaching assistants. In some cases, these are an underused resource, and so using them to give individual or small-group speaking practice for these exams is a way of gaining added value. Third, the availability of appropriate listening material with a range of different voices made it possible to give students more practice at what had been a rather neglected skill, since most of the listening input had previously come from the teachers themselves.

Finally, the thorny question as to whether using *Cambridge English: Young Learners* meant more work for an already overworked group of professionals received an answer which was surprisingly positive. Although sometimes teachers had to put in extra time making copies of exam papers and correcting them, they did not regard this as particularly onerous, and the effort was well compensated by the wide availability of suitable material, and the ease of correcting exams. The part of the test which was the most difficult to prepare for in these schools was Speaking: although some schools had teaching assistants, many did not, and had to draw on the occasional participation of student teachers, or else simply practise the speaking in whole-group sessions.

In general, the story which this article tells is one of positive integration of *Cambridge English: Young Learners* into the primary curricula in schools mainly in the *concertado* sector in Navarra. These exams match well with existing good practices, and encourage teachers to move towards a more balanced curriculum which gives equal weight to the four skills. Teachers in this study greatly appreciate the exams' clarity of criteria, transparency of exam content and format, and balance of skills. They report an increase in their students' motivation to learn, as well as a parallel enhancement of their own motivation to teach. The availability of practice material is evidently a key factor in encouraging teachers to respond positively to the idea of preparing their students for the exams. Nonetheless, our findings are limited by the fact that our sample only included teachers who were already involved in preparing students for the exams, and was dominated by teachers from the *concertado* sector. Indications of resistance from colleagues, particularly in state schools, shed some light on the situation that may exist outside the bounds of this sample. Reports from state primary schools indicate that this sector is particularly challenging terrain for international examining boards, for a variety of reasons. Further qualitative research of the kind carried out in the present study is needed in order to gain a broader and deeper understanding of the attitudes of Spanish educational professionals towards Cambridge English examinations.

## References

Ashton, K, Salamoura, A and Diaz, E (2012) The BEDA impact project: A preliminary investigation of a bilingual programme in Spain, *Research Notes* 50, 34–42.

Blanco, J and Nicholson, D (2010) Cambridge ESOL and Spanish school networks, *Research Notes* 40, 9–13.

Chambers, L, Elliott, M and Jianguo, H (2012) The Hebei Impact Project: A study into the impact of Cambridge English exams in the state sector in Hebei province, China, *Research Notes* 50, 20–23.

Cheng, L Y (2005) *Changing Language Teaching through Language Testing: A Washback Study*, Studies in Language Testing volume 21, Cambridge: UCLES/Cambridge University Press.

Gobierno de Navarra (2013) *Modelos de Evaluación* (4º curso de EP), available online: www.educacion.navarra.es/web/dpto/primaria-por-asignaturas

Gu, X, Khalifa, H, Yan, Q and Tian, J (2012) A small-scale pilot study investigating the impact of *Cambridge English: Young Learners* in China, *Research Notes* 50, 42–47.

Khalifa, H, Nguyen, T and Walker, C (2012) An investigation into the effect of intensive language provision and external assessment in primary education in Ho Chi Minh City, Vietnam, *Research Notes* 50, 8–19.

Salamoura, A, Hamilton, M and Octor, V (2012) An initial investigation of the introduction of Cambridge English examinations in Mission laïque française schools, *Research Notes* 50, 24–31.

Saville, N (2012) Applying a model for investigating the impact of language assessment within educational contexts: The Cambridge ESOL approach, *Research Notes* 50, 4–8.

Spanish Eurydice Unit (2009) *Structures of Education and Training Systems in Europe: Spain*, available online: eacea.ec.europa.eu/education/eurydice/documents/eurybase/structures/041_es_en.pdf

## Appendix 1: Interview schedule

1 Reasons why the school has decided to incorporate Cambridge English: Young Learners

How long has your school been preparing students for *Cambridge English: Young Learners*?
How was the decision to start preparing students for *Cambridge English: Young Learners* made?
Were you involved in this decision?
What were the main reasons for deciding to offer *Cambridge English: Young Learners* preparation to the students in your school?

2 Teachers' attitudes to adopting *Cambridge English: Young Learners*

How did you feel when this decision was made?
Do you think it was a good decision?
In your opinion, what could be the advantages and disadvantages of preparing for *Cambridge English: Young Learners* . . .
. . . for the school,
. . . for the teacher,
. . . for the students,
. . . for the students' parents?
Do you think the *Cambridge English: Young Learners* exam is a useful measure of your learners' English ability? Why (not)?
How important is it to you that your students do well on the exam?

How do you think your students feel about taking *Cambridge English: Young Learners*?

3 Washback effects of *Cambridge English: Young Learners* on classroom practice

What do you think is the most effective method to teach English? Why?
What kind of classroom activities do you think are best for improving your students' level of English?
How did you learn English? Did you think it was a good method?
How often do you attend training sessions?
Did you receive any training sessions about the *Cambridge English: Young Learners* exams?
Which materials do you use in your English lessons? (textbook, past exam papers, other . . .)
Have you changed anything about the materials you use in class since your school implemented *Cambridge English: Young Learners*?
How important do you think the following aspects of English are for your students? Why? How much time do you spend on teaching the following aspects or skills?
Grammar
Vocabulary
Reading
Writing
Listening
Speaking
How much time did you use to spend on them before your school implemented *Cambridge English: Young Learners*? Is there a difference?
How often do you use the following teaching activities in class?
Whole class activities
Individual work
Pair work
Group work
Has the implementation of the *Cambridge English: Young Learners* exam changed anything about the kinds of activities you do in class or the amount of time you spend on certain activities?
How often do you use English in class with your students?
Do your students use English in class to talk to you or to talk to other students?
Has the amount of English used by you or your students changed since your school implemented *Cambridge English: Young Learners*?
Has the decision to prepare the students for *Cambridge English: Young Learners* given you any extra work?
Do you plan your lessons in a different way now?
Do you find it easy or difficult to prepare your students for *Cambridge English: Young Learners*? What difficulties do you have?

## Appendix 2: Background information and results

**Table 1: Teachers' experience and qualifications**

| Teacher | Sex | Age | L1 | Teaching experience | Experience with exams | Degrees |
|---|---|---|---|---|---|---|
| **C1 Ann** | F | 32 | English | 8 years | Many years at same school | BA, MA |
| **C1 Marie** | F | 57 | Spanish/French | 29 years | Many years at same school | Magisterio* |
| **C2 Maria** | F | 38 | Spanish/English | 19 years | Two years at same school | Magisterio |
| **C2 Jose** | M | 22 | Spanish | A few months | No previous experience | Magisterio |
| **C3 Antonio** | M | 40 | Spanish/Basque | 20 years | 7 years at same school | Magisterio |
| **C3 Rodrigo** | M | 26 | Spanish | 4 years | 4 years at same school | Licenciatura** and Magisterio |
| **C3 Jesus** | M | 47 | Spanish | 20+ years | Many years at same school | Licenciatura and Magisterio |
| **C4 Irantzu** | F | 30 | Spanish | 7 years | 4 years at same school | Magisterio |
| **C4 Elena** | F | 46 | Spanish | 16 years | 4 years at same school | Magisterio |
| **C4 Rocio** | F | 29 | Spanish | 9 years | 4 years at same school | Magisterio and Kindergarten |
| **C4 Cristina** | F | 28 | Spanish | 6 years | 4 years at same school | Magisterio |
| **C5 Rosy** | F | 48 | English | 20 years | 4 years at same school | BA, Magisterio |
| **C6 Jack** | M | 52 | English | 27 years | Many years at same school | BA, Magisterio |
| **C6 Marta** | F | 36 | Spanish | 10 years | Many years at same school | Licenciatura and Magisterio |
| **C7 Asier** | M | 35 | Spanish/Basque | 12 years | In previous school | Licenciatura and Magisterio |
| **C8 Pilar** | F | 46 | Spanish | 9 years | 7 years at same school | Magisterio |
| **C8 Eva** | F | 40 | Spanish | 15 years | 7 years at same school | Licenciatura and Magisterio |
| **C9 Jorge** | M | 30 | Spanish | 7 years | 6 years at same school | Magisterio |
| **C10 Helen** | F | 52 | English | 30 years | Many years at same school | BA, PGCE, Magisterio |
| **P1 Chloe** | F | 41 | English | 17 years | In previous school | BA, Magisterio |
| **P2 Ignacio** | M | 36 | Spanish | 7 years | In previous school | Licenciatura and Magisterio |
| **P3 Fernando** | M | 42 | Spanish | 19 years | In previous school | Licenciatura and Magisterio |

*Magisterio is the Spanish Primary Teaching Degree.

**Licenciatura is the Spanish equivalent of BA or BSc.

**Table 2: Hours of English/Content and Language Integrated Learning (CLIL)/exam preparation per week**

| School | Hours of English/CLIL per week | Hours spent on exam preparation |
|---|---|---|
| **San Cernin** | 4 English/3 CLIL | 1 per week in second/third term of 6th year of primary |
| **La Compasión** | 5 English/2 CLIL | 1 hour per week |
| **El Redín-Miravalles** | 7 English/4 CLIL | Integrated into all English classes |
| **Liceo Monjardín** | 4 English/3 CLIL | 1 hour per week in 6th year of primary |
| **Escolapios** | 4 English/5 CLIL | 1 hour per week in second/third term of 6th year of primary |
| **Sagrado Corazón** | 5 English | 1 hour every 2 weeks in 6th year of primary |
| **Regina Pacis** | 4 English/1 CLIL | Starting this year |
| **San Ignacio** | 4 English (6th year) 3 English/2 CLIL (other years) | Mock exam once a term |
| **Irabia-Izaga** | 4 English/3 CLIL | 1 hour per week in second/third term |
| **Nuestra Señora del Huerto** | 4 English/8 CLIL | 1 hour per week |
| **CP Catalina de Foix** | 5 English/CLIL varies according to year | No preparation |
| **CP Santa Maria Los Arcos** | 5 English | 1 hour per week |
| **CP Ermitagaña** | 4–5 English | Starting this year |

## Appendix 3: Tables of themes by school

**Table 3: Advantages of preparing students for Cambridge English: Young Learners**

| | School's image | External validation | Preparing for future exams | Students' motivation | Helps students improve | Teachers' motivation | Framework/ aim for teachers | Information on weak areas | Different way of learning/ teaching | Good format/ material | Includes oral |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **C1A** | X | X | | | | | X | | X | | |
| **C1B** | X | X | X | X | X | X | | | | | |
| **C2A** | X | | X | | | | | | X | X | |
| **C2B** | | X | X | | | | X | | | | |
| **C3A** | X | | X | | | X | | | | | |
| **C3B** | X | | | X | | X | | | | | |
| **C3C** | X | X | | X | | X | X | X | | X | X |
| **C4A** | | X | X | X | X | X | | X | X | X | |
| **C4B** | X | X | X | X | | | X | X | | X | |
| **C5** | | | | X | | | X | | | | X |
| **C6** | | | X | X | X | | | | X | X | X |
| **C7** | X | X | X | X | X | X | | | X | X | |
| **C8** | X | X | X | X | X | | | X | | | X |
| **C9** | | X | | X | X | X | | X | | X | |
| **C10** | | X | X | X | X | X | X | | | X | X |
| **P1** | X | X | X | | | X | | | | X | |
| **P2** | | X | | X | X | X | | | X | X | |
| **P3** | X | X | | | | X | | | | X | |

**Table 4: Disadvantages of preparing students for Cambridge English: Young Learners**

| | Money | Nervous students | Weak students | Too young | Low levels | Pressure on teachers | Extra work |
|---|---|---|---|---|---|---|---|
| **C1A** | X | | | | | | |
| **C1B** | X | X | | | | | |
| **C2A** | | | | | X | | X |
| **C2B** | | | X | | X | X | |
| **C3A** | | | | X | | | |
| **C3B** | X | X | | X | | X | |
| **C3C** | X | | X | | | | |
| **C4A** | X | X | | X | | | X |
| **C4B** | X | X | | X | | | |
| **C5** | X | | | X | | | |
| **C6** | X | | | X | | | |
| **C7** | X | | | X | | X | |
| **C8** | | X | X | X | | | |

**Table 4: (continued)**

| | Money | Nervous students | Weak students | Too young | Low levels | Pressure on teachers | Extra work |
|---|---|---|---|---|---|---|---|
| **C9** | | | | X | | | |
| **C10** | X | | | X | X | | X |
| **P1** | | X | | | | | |
| **P2** | X | X | | X | | | |
| **P3** | X | | | | | | X |

**Table 5: Washback of Cambridge English: Young Learners**

| | Vocabulary | Focus on speaking | Mock exams | Using Cambridge English for internal assessment | Focus on spelling/ writing | Change course book | Using Cambridge English materials in regular class | Speaking with assistant |
|---|---|---|---|---|---|---|---|---|
| **C1A** | | X | X | | X | X | | |
| **C1B** | X | X | X | X | X | | | |
| **C2A** | X | | X | | | | | |
| **C2B** | | | X | | | | | |
| **C3A** | | | X | | X | | | |
| **C3B** | X | | X | | | X | | |
| **C3C** | X | X | X | X | X | | | X |
| **C4A** | X | X | X | | | | | |
| **C4B** | X | | | X | X | | X | |
| **C5** | X | | X | X | | X | | X |
| **C6** | X | X | X | X | | | | |
| **C7** | X | | X | | X | | | |
| **C8** | X | | X | | X | X | | X |
| **C9** | | X | X | X | | | | |
| **C10** | X | X | X | X | | X | X | |
| **P1** | X | | | | X | | | |
| **P2** | | X | X | X | X | X | | |
| **P3** | X | X | | | X | | X | |

# The impact of TKT on Chinese teachers' teaching beliefs, knowledge and practice

**LIYAN HUANG** GUANGDONG ACADEMY OF EDUCATION, CHINA
**ANGELO PAPAKOSMAS** GUANGZHOU EDUCATION BUREAU, CHINA

## Introduction

The study described in this paper is a funded project by Cambridge English Language Assessment, and examines the impact of the Cambridge English *Teaching Knowledge Test* (*TKT*) on Chinese teachers' teaching beliefs, knowledge and practice. Specifically, two research questions were addressed. Firstly, what is the impact of *TKT* on Chinese teachers' teaching beliefs, knowledge and practice? We attempted to explore the scope and intensity (Cheng 2005) of *TKT* impact. Secondly, what are the roles of contextual factors in shaping *TKT* impact? We investigated the variation of the perceived impact of age, teaching experience, education level, education sector, way of preparing for *TKT, TKT* preparation duration, and time for sitting *TKT*. We also examined how the contextual factors including teachers' beliefs and attitudes, the influence of teachers' important referents, and teachers' perceptions of their ability to carry out a given action, have influenced *TKT* impact.

Modules 1 to 3 of *TKT* were investigated in the present study. Module 1 tests the knowledge of the terms and concepts of English language teaching and it focuses on the factors underpinning the learning of English. Module 2 tests the knowledge and skills necessary for lesson planning. This module also covers knowledge about assessment and resource use. Module 3 tests knowledge of what happens in the classroom during language learning, the teacher's role in classroom management as well as methods used to manage and make the most of interactions in the classroom[1].

## Impact

Impact is defined as 'any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole' (Wall 1997:291). The scope of impact can be examined from various perspectives, ranging from individual stakeholders (teachers, students, test designers, material developers etc.) to systemic impact (on education systems, curriculum development, social equality). Given such variation, even when a single group of stakeholders (e.g. teachers) is investigated, operationalising impact for research purposes is far from straightforward. While different studies (e.g. Huang 2011, Wall 2005) have examined the impact of tests of language proficiency on teachers, these have overwhelmingly been qualitative in nature, Valazza (2008) being a notable exception).

When attempting to measure impact more quantitatively, as is done in the present study, the dimensions examined must be precisely defined, then adequate metrics or variables identified to effectively operationalise those dimensions.

The *TKT* dimensions used here are based on those defined by Grossman (1990, adapted from Spratt 2004). According to Spratt (2004:2–3), *TKT* aims to test teaching knowledge rather than teaching ability or performance in classroom teaching. The four components of teaching knowledge are outlined below.

*Subject matter knowledge*. This concerns the understanding of the facts, concepts and terminology of a subject discipline. This requires the knowledge of the English language system (phonology, lexis, grammar and function/discourse), language processing and production, and four language skills including listening, speaking, reading and writing, and their sub-skills.

*General pedagogic knowledge*. This can be defined as the knowledge of general principles of teaching and learning which are applicable across subjects. For *TKT* this involves management of resources (accessibility, adequacy, appropriacy, authenticity and variety, etc.) and management of learning (motivation, involvement, organising of learning, affective dimension and learner empowerment, etc.).

*Pedagogic content knowledge*. This concerns the representation of the subject matter through examples, analogies and procedure, to make it comprehensible to students. For *TKT* this involves language learning strategies (risk taking, self-monitoring, tolerance of ambiguity, etc.) and language processing and production (four language skills – reading, writing etc. and their sub-skills). This may also be categorised as part of *subject matter knowledge*.

*Contextual knowledge*. This can be defined as the knowledge of educational aims, students and other contextual factors, which would inform the application of the other three types of knowledge. Broadly, contextual knowledge includes the school the teacher works in, the specific set of students they teach and the wider educational context in which they operate.

The above-mentioned components of teaching knowledge were used to guide the present research into the impact of *TKT* on teachers.

To explore and explain why impact has taken the form it has, the factors including those derived from demographic data and other contextual factors that potentially shape the response of teachers to a test need to be examined. The Theory of Planned Behaviour (TPB) (Ajzen 1985, 1991, 2006) was developed to explain human social behaviour and it identifies three major variables which influence human action:

- *Attitude*: The extent to which a person believes a given behaviour will result in a beneficial outcome. A person's attitude to a given behaviour is most immediately influenced by their *behavioural beliefs*. The more positive an attitude towards a behaviour, the greater a person's inclination to exhibit that behaviour.
- *Subjective norm*: People's perceptions of social pressure to perform or not to perform a behaviour. The *subjective norm* is determined by *normative beliefs*, the perceived behavioural expectations of important referent individuals or groups.
- *Perceived behavioural control*: People's perceptions of their ability to perform a given behaviour. It is determined by their *control beliefs* about the power of each facilitating or impeding factor.

The three above-mentioned variables are considered as the contextual factors that have an influence on teachers' teaching decisions (Huang 2011); in the present research, the decision is 'whether or not to put what we have learned from *TKT* into teaching practice'. We decided that TPB could be well suited to analyse *TKT* impact by examining teachers' attitudes towards applying *TKT* in practice, the social pressures exerted on them to do so, and their perceptions of their ability to apply *TKT*. Therefore, TPB was used to inform the questionnaire design aiming to explain why teachers have or have not put what they have learned from *TKT* into teaching practice.

## Research method

The research presented here can be characterised as a mixed methods study with exploratory design (Creswell and Plano Clark 2011). This approach was chosen so that both qualitative and quantitative data could be use to analyse and interpret the impact of *TKT*. Divided into two stages, the first, qualitative stage (using focus group interviews), which took place in May 2013, was used to get a holistic impression of teachers' views on *TKT* impact and to assist in the design of the questionnaire used in the second stage. Stage 2 began with a pilot study in November 2013 to improve the questionnaire design and delivery. Following the pilot, and a refinement of the questionnaire, we contacted past *TKT* takers from around China using a web-based survey to investigate the impact of *TKT* on Chinese teachers' teaching beliefs, knowledge and practice.

[1] Taken from *TKT* handbook for teachers: www.cambridgeenglish.org/images/22188-tkt-kal-handbook.pdf

## Participants

There was a total of 20 primary school teachers in the focus groups, consisting of teachers of English from Guangzhou, China, who were sent to Britain by Guangzhou Education Bureau in 2011 for professional teacher training, which resulted in their taking the first three modules of *TKT*.

As for the survey respondents, with the help of Cambridge English Language Assessment[2] and the present researchers' networks, we invited approximately 230 past *TKT* takers to participate in the survey in December 2013 and received 206 valid responses, including 83 from Hebei province, 55 from Guangdong province, 47 from Beijing, one from Xinjiang province and 20 from other provinces and cities. The participants cover the eastern, southern, western and northern part of China, which increases the representativeness of the sample of the target population, i.e. Chinese *TKT* takers. The majority of the respondents were university graduates, below 40, well experienced, teaching at primary and secondary school. More than half of them were from private schools. About half of them participated in training courses before sitting *TKT*. Table 1 illustrates the general characteristics of the respondents.[3]

## Instruments

Focus groups were conducted to elicit the participants' general ideas about *TKT* impact on their understanding of English teaching, their knowledge and their teaching practice. Prior to the interviews, some questions similar to the research questions were prepared to guide the discussion, which was carried out in Chinese to ensure complete understanding and clear expression by the participants.

The findings derived from focus groups, the *TKT* construct defined by Spratt (2004) and the *TKT* teacher questionnaire designed by Valazza (2008) assisted in the present questionnaire design. The questionnaire (see Appendix 1) included a total of 46 items, divided into three sections. Section One comprised eight questions eliciting respondents' demographic characteristics and temporal aspects of their *TKT* experience, such as how long they prepared for *TKT* and when they took *TKT* (items 1–8). Twenty-eight Likert items (items 9–36), consisting of a 5-point response scale (Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree), made up Section Two of the questionnaire. These specifically targeted the four dimensions of teaching knowledge that are covered in *TKT*. Section Three of the questionnaire had 10 Likert items (37–46) representing the three dimensions of Ajzen's TPB, i.e. behavioural beliefs and attitudes, normative beliefs and subjective norms, control beliefs and perceived behavioural controls.

## Data analysis

Content analysis was used in focus group data analysis and we mainly reviewed the transcripts and identified the salient issues.

**Table 1: Respondents' demographic characteristics**

| Item | Options | Count | Percentage |
|---|---|---|---|
| **Age** | Below 40 | 171 | 83% |
| | Between 41 and 50 | 30 | 15% |
| | Over 50 | 5 | 2% |
| **Teaching experience** | Less than 5 years | 44 | 21% |
| | 5 to 10 years | 54 | 26% |
| | More than 10 years | 108 | 53% |
| **Education level** | Post-secondary education | 12 | 6% |
| | College degree | 165 | 80% |
| | Master degree | 29 | 14% |
| **Education sector** | Primary school | 89 | 43% |
| | Secondary school | 91 | 44% |
| | University | 26 | 13% |
| | Private school | 138 | 67% |
| | Public school | 68 | 33% |
| **Way of preparing for *TKT*** | Taking course in China | 97 | 47% |
| | Taking course abroad | 38 | 19% |
| | Taking course at home and aboard | 15 | 7% |
| | Self-study | 56 | 27% |
| **Preparation duration** | 1–3 months | 146 | 71% |
| | 6 months or more | 60 | 29% |
| **Time of sitting *TKT*** | Before 2010 | 58 | 28% |
| | Since 2010 | 148 | 72% |

Various statistical methods were used to examine the questionnaire data. As for the second section of the questionnaire, which is concerned with *TKT* impact on teachers' knowledge, Cronbach's alpha was used to measure the four dependent variables' (subject matter knowledge, general pedagogic knowledge, pedagogic content knowledge, contextual knowledge) internal consistency or reliability, i.e. the extent to which the individual items together represent the underlying construct. Additionally, the analytical method used to calculate the perceived positive impact of *TKT* on these variables was the sum total of positive responses to each item. In other words, all teachers who answered either 'Agree' or 'Strongly agree' to an item were added together to calculate a single figure of agreement. The items were then ranked based on the percentage of positive responses to determine the relative strength of *TKT* impact.

Analysis of variance was then carried out between each of the *TKT* impact dependent variables and the seven independent variables (age, teaching experience, education level, education sector, way of preparing *TKT*, *TKT* preparation duration, and time of sitting *TKT*) from Section One of the questionnaire. Variance between groups was analysed using both one-way ANOVA (which assumes the data is scalar and bases calculation on group means) and Kruskal-Wallis (used for ordinal data and based on group ranked medians). Where significant differences were identified, post hoc tests were used to identify which paired groups varied. For ANOVA, post hoc tests were Tukey and Games Howel (the latter used when Levene test showed data did not satisfy the ANOVA assumption of equal

[2] Special thanks go to Su Shanshan from the Cambridge English Language Assessment China Office.

[3] Because of a small sample size, 51+ age group was combined with 41–50 age group into a new group of 41+.

distribution of variance) and for Kruskal-Wallis, the Mann-Whitney U test was applied. All analyses were carried out using Statistical Package for the Social Sciences (SPSS) v.20.

## Results and discussion

In this section, the findings of the focus groups and questionnaire are presented and discussed.

### Focus groups

In the focus group sessions, it was suggested by the teachers that the *TKT* and its preparation training programme had considerable positive influence on their teaching beliefs, knowledge and behaviour. The areas on which *TKT* had positive influence included teaching resource use, teaching methods and skills, teachers' language development, planning of lessons, etc. Quite a few teachers recognised the impact of *TKT* (preparation) on teaching resource use; for example, the participating teachers noted the value, in principle at least, of a shift away from the traditional Chinese classroom reliance on the textbook. In order to stimulate students' interest, a wider array of language inputs were required. The teachers also reported that the knowledge they gained through *TKT* and the resulting shift in some of their teaching beliefs translated into changes in their actual classroom behaviour. For example, some teachers claimed a shift in the type of interaction with students, with efforts to move away from the traditions of rote learning and mechanical drills to more interpretive and interactive exchanges, as well as the use of more co-operative classroom work modes such as pair work or group work. Also, the skills acquired and used to plan lessons were discussed. Teachers repeatedly praised the knowledge gained in this field and cited ways in which they were able to put this knowledge to practical use. Another positive impact of *TKT* raised by teachers was its effect on their own language development. Most teachers stated that their speaking and listening skills improved considerably after sitting *TKT*.

However, teachers also referred to a number of obstacles or limitations to their capacity to transfer some of what they had learned into their own classroom teaching, such as resistance from school leaders and colleagues, pressure from examinations and parents, lack of school equipment, large class size, limited teaching time and absence of teaching resources.

The findings derived from focus group interviews informed us of some of the impact areas of *TKT* on beliefs, knowledge and actions, such as teaching resource and teaching methods and skills, which provides part of the answer to the first research question (namely, what is the impact of *TKT* on Chinese teachers' teaching beliefs, knowledge and practice?) and assisted in the questionnaire design. In addition, the findings indicated some contextual factors such as how colleagues' perceptions towards *TKT* and restriction of school facilities might impede the intensity of *TKT* impact.

### Questionnaire

The results of the questionnaire data analysis were used to answer both the first and the second research questions. In this section, firstly, findings about *TKT* impact on teachers, which derived from Section Two of the questionnaire, will be presented. Secondly, we will describe the roles of contextual factors in shaping *TKT* impact based on the findings from both Sections Two and Three. Before turning to the findings of *TKT* impact on teachers, let us present the findings concerning the internal consistency of the questionnaire.

The Cronbach's alpha measure for each of the *TKT* impact variables used here was:

- subject matter knowledge (items 9 to 15) – Cronbach's alpha = 0.952
- general pedagogic knowledge (items 16 to 25) – Cronbach's alpha = 0.95
- pedagogic content knowledge (items 26 to 33) – Cronbach's alpha = 0.954
- context knowledge (items 34–36) – Cronbach's alpha = 0.89.

With a reliability coefficient of 0.70 or higher considered an acceptable level of internal consistency in most social science research, each of the constructed variables had a high level of internal reliability indicating they adequately represented the construct. In addition, factor analysis showed that the relationship of the individual items within each of the *TKT* variables was unidimensional.

*Impact of TKT on teachers*

Section Two of the questionnaire was made up of four areas, consisting of 28 items representing the *TKT* impact on teachers' knowledge. These were used to answer the first research question. The findings are presented in Table 2.

Overall the teachers appear to assess the impact of participating in *TKT* as overwhelmingly positive, with the collective positive contribution made by *TKT* to improvement in the individual items ranging from 83% to 63%. One feature that stands out is the relative concentration of items that refer specifically to teaching abilities and changes in actual teaching behaviour, either inside the classroom (e.g. more varied use of teaching materials) or outside the classroom (e.g. improved lesson planning). Amongst the nine items with a total positive response of 75% or more, six related specifically to abilities and actions. These included improved teaching skills (item 26, 80% positive response), the use of a greater variety of classroom activities (item 23, 78% positive response), improved planning (items 21 and 22), material evaluation (item 20) and more effective use of teaching resources (item 36, 75% positive). Additionally two others, relating to learner assessment (item 33) and student needs identification (item 35), had a total positive response measure of 69% or more. This may suggest that, although *TKT* claims not to measure teaching ability, teachers did perceive it as having a very direct positive influence on their overall teaching abilities, which resulted in changes in what they did both inside and outside the classroom. This would appear to indicate that the teachers were able to translate their gains in knowledge to actual changes in their behaviour.

The teachers' responses also indicated a substantial improvement in areas related to teaching theory, with a better understanding of teaching methods (item 16) and teaching aims (item 34) taking the top two positions; improved understanding of teaching theory (item 17) also featured prominently (78% positive response).

**Table 2: Impact of TKT on teachers**

| Impact area | Items | Strongly agree | | Agree | | Total positive response | |
|---|---|---|---|---|---|---|---|
| | | Count | Percentage | Count | Percentage | Count | Percentage |
| **Subject matter knowledge** | 14. reading | 54 | 26% | 96 | 47% | 150 | 73% |
| | 10. vocabulary | 50 | 24% | 91 | 44% | 141 | 68% |
| | 13. speaking | 50 | 24% | 88 | 43% | 138 | 67% |
| | 12. listening | 50 | 24% | 86 | 42% | 136 | 66% |
| | 9. grammar | 54 | 26% | 77 | 37% | 131 | 64% |
| | 15. writing | 40 | 19% | 90 | 44% | 130 | 63% |
| | 11. phonetics | 46 | 22% | 83 | 40% | 129 | 63% |
| **General pedagogic knowledge** | 16. teaching methods | 87 | 42% | 84 | 41% | 171 | 83% |
| | 23. activity variety | 69 | 34% | 92 | 45% | 161 | 78% |
| | 17. teaching theory | 93 | 45% | 67 | 33% | 160 | 78% |
| | 21. plan for students' needs | 65 | 32% | 94 | 46% | 159 | 77% |
| | 22. plan lessons | 73 | 35% | 86 | 42% | 159 | 77% |
| | 20. material evaluation | 70 | 34% | 87 | 42% | 157 | 76% |
| | 18. teaching self-aware positive | 63 | 31% | 89 | 43% | 152 | 74% |
| | 19. teaching self-aware negative | 59 | 29% | 90 | 44% | 149 | 72% |
| | 25. learner style awareness | 51 | 25% | 94 | 46% | 145 | 70% |
| | 24. student enjoyment | 45 | 22% | 92 | 45% | 137 | 67% |
| **Pedagogic content knowledge** | 26. teaching skills | 90 | 44% | 74 | 36% | 164 | 80% |
| | 33. learner assessment | 50 | 24% | 101 | 49% | 151 | 73% |
| | 29. reading strategies | 50 | 24% | 99 | 47% | 149 | 71% |
| | 27. listening strategies | 48 | 23% | 94 | 46% | 142 | 69% |
| | 28. speaking strategies | 45 | 22% | 94 | 46% | 139 | 68% |
| | 32. grammar teaching | 40 | 19% | 96 | 47% | 136 | 66% |
| | 30. writing strategies | 43 | 21% | 90 | 44% | 133 | 65% |
| | 31. vocabulary teaching | 48 | 23% | 83 | 40% | 131 | 64% |
| **Contextual knowledge** | 34. teaching aims | 79 | 38% | 91 | 44% | 170 | 83% |
| | 36. teaching resource use | 62 | 30% | 92 | 45% | 154 | 75% |
| | 35. student needs identification | 56 | 27% | 86 | 42% | 142 | 69% |

At the other end of the scale, it is also evident that teachers felt that *TKT* had far less influence or effect on those areas we have categorised as related to subject matter knowledge, although even in these areas the effect was still clearly positive. Items related to the language system (items 9 to 11), and to language processing and production (items 12 to 15) were all ranked in the bottom half with positive sum totals between 63% and 68%. The exception to this was item 14, related to an improved understanding of reading skills and sub-skills, which had a positive sum value of 73%.

This relatively low level of impact on subject matter knowledge may in part be a reflection of the language education and training they receive. Both, when learning English as middle school students and in their subsequent college and university education, the focus of the instruction they receive is overwhelmingly grammar and skills oriented. It may be the case that the foundation of the Chinese teachers' knowledge in these areas is already well developed and therefore there is less capacity for improvement. By contrast, teaching theories, and aspects of pedagogical knowledge, particularly those related to more contemporary communicative and learner-oriented teaching methods, are less well covered in their university studies and consequently these are more prominently identified as areas where teachers perceived greatest benefits.

### Variation in TKT impact

This section attempts to determine whether there was any significant variation in *TKT* impact amongst respondent groups based on age, teaching experience, education level, education sector, way of preparing *TKT, TKT* preparation duration, and time of sitting *TKT*. The result of analysis based on 'Region of China' is not included because the findings are not statistically significant.

The findings using both ANOVA and Kruskal-Wallis (and their respective post hoc tests) were broadly similar, so only the ANOVA findings are reported here. There was no statistically significant difference in variance in impact based on age[4] and teaching experience, so only the variables of education level, education sector, way of preparing *TKT, TKT* preparation duration, and time of sitting *TKT* will be discussed further.

*Education level and TKT impact*

It was found that the *TKT* impact on teachers' pedagogic content knowledge differed significantly as a function of education level, F (2, 202) = 3.976, p = .02. Tukey post hoc

[4] Because of a small sample size, the 51+ age group was combined with the 41–50 age group into a new group of 41+.

comparisons between the groups indicated that the group with a Master's degree education level (M = 3.5, 95% CI [3.26, 4.58]) had a significantly lower pedagogic content knowledge mean than the undergraduate degree group (M = 3.93, 95% CI [3.82, 4.04]), p = .015. Comparisons between post-secondary and the other two groups were not statistically significant at p<.05.

The finding of lower impact for the teachers with a Master's degree compared to their counterparts with an undergraduate degree may indicate that the additional knowledge and skills they acquired during postgraduate studies meant that some of the *TKT* content was already familiar to them. For those with undergraduate qualifications, who have less exposure to such pedagogic concepts and ideas while students, there was more scope of acquisition of new knowledge and consequently greater positive impact from *TKT*.

*Education sector and TKT impact*

The results of one-way ANOVA analysis showed there was a statistically significant effect of teachers' school sector on:

- the *TKT*'s subject matter knowledge impact F (6, 199) = 2.562, p = .02
- the *TKT*'s pedagogic content knowledge impact F (6, 199) = 2.66, p = .017
- the *TKT*'s context knowledge impact F (6, 199) = 2.6, p= .019.

For subject matter knowledge impact, the Tukey post hoc test showed that the only statistically significant variation of mean results was between public sector university teachers (M = 3.3, 95% CI [2.92, 3.68]) and public sector elementary school teachers (M = 4.1, 95% CI [3.97, 4.35]) p = .013. The post hoc Kruskal-Wallis test, and Mann-Whitney U test, also identified significant variance (of medians) between these two groups (at p<.01). In addition it also found that ranked median scores for public sector university teachers were statistically significantly lower than those of public sector senior and junior middle school teachers.

For both pedagogic content knowledge impact and for context knowledge impact, Tukey identified only one pair-wise significant difference, that between public sector university teachers and their public sector elementary counterparts. In both cases university teacher mean scores were found to be statistically significantly lower than public school elementary teachers at p <.05. With the Mann-Whitney U test, for pedagogic content knowledge, university teacher ranked median scores were found to be significantly lower/different from those of public senior middle school, junior middle school, and elementary school teachers. For context knowledge impact, the non-parametric test found significant differences between public sector university teachers and both public junior and public elementary teachers. All significant pair-wise differences with the Mann-Whitney U test were at p<.01.

In terms of interpretation, once again the trend of statistically significant variance in impact, in the public sector at least, indicates that teachers at the lower end seem to derive more benefit from *TKT*. In the Chinese educational context the general perception is that teachers' competence (as well as the training they have) is reflected in their location in the education sector hierarchy. Thus, for example, the most able teachers in the secondary sector are typically assigned by school leaders to teach senior classes. As with the difference in education level, the net benefit to teachers of *TKT* was in inverse relationship to their position in the educational hierarchy and may reflect the difference in pre- and post-placement training they receive. Whatever the explanation for the difference, the data revealed that *TKT* had stronger impact on the teachers at the lower end of the educational hierarchy.

*Way of preparing for TKT and TKT impact*

Four *TKT* preparatory course types (independent study; formal training course in China; formal training course abroad; formal training course partly in China, partly abroad) were examined to see if they made any contribution to differences in the impact of the *TKT* on teachers (see item 6 of the questionnaire for details). The analysis found that course type had a statistically significant effect on:

- the *TKT*'s *subject matter knowledge* impact F (3, 202) = 4.84, p=.00
- the *TKT*'s *general pedagogic knowledge* impact F (3, 202) = 3.30, p = .022; and
- the *TKT*'s *context knowledge* impact F (3, 202) = 4.02, p= .008.

Because the Levene test indicated a lack of normal variance between groups the Games Howell was used as the post hoc test which found the following:

- For subject matter knowledge – the impact on Group 1 teachers (formal training course in China) was significantly lower than that of Group 3 (formal training in China and abroad, p=.027) and Group 2 (formal training abroad only, p=.004). For teachers who studied independently in China the subject matter impact was also significantly lower than for Group 2 (p=.009) and Group 3 (p=.026). But there was no significant variance between those who prepared entirely in China.
- For general pedagogic knowledge impact mean scores only varied statistically significantly between self-study teachers and those who took part in the *TKT* programme entirely abroad (p = .041).
- For context knowledge there were significant differences between Group 1 and Group 3 teachers (p = .039) while Group 4 teachers had lower impact scores than Group 2 (p = .019) and Group 3 teachers (p = .009).

The findings seem to suggest that improvement was greatest for the teachers who took part in training abroad, either entirely or partially. This may be partly attributable to the additional exposure teachers get to language and overseas teaching methods through immersion abroad. It may also suggest that current *TKT* formal preparatory training programmes in China are not as well developed or as effective as those abroad, which would in part explain the lack of difference between Groups 1 and 4.

*TKT preparation duration and TKT impact*

The original ANOVA and Kruskal-Wallis analyses found a significant difference in mean/median scores for all four impact variables based on course duration. However post hoc tests showed the only significant variation (for all four dependent variables) was between the group who had taken over a year to complete the course and all other groups. It was therefore decided to recode the data and create two course duration groups, those completing the course in one year or less, and those taking over one year. An independent sample *t*-test was then used to measure the variance of means. The *t*-test reconfirmed that teachers who took longer than one year to complete the test were found to have reported the impact of *TKT* across all four dimensions to be significantly lower than for teachers who had completed the *TKT* course and tests (either through self-study or in an organised course) in one year or less.

*Time for sitting TKT and TKT impact*

The one-way ANOVA test found significant differences in pedagogic content knowledge impact and general pedagogic knowledge impact based on the year teachers completed their final *TKT* module, although the Kruskal-Wallis test found no such significant variation. The Tukey test on general pedagogic knowledge impact for this variable found only one significant difference, that between teachers completing *TKT* before 2010 and those who completed *TKT* in 2011 (p = .033). For pedagogic content knowledge (which failed the Levene test for homogeneity of variance), the Games Howell test found a difference only between those completing *TKT* in 2011 and 2013 (p = .027). When the data was re-examined to see if a difference existed between those completing *TKT* before 2012 and those completing it in 2012 and after, an independent samples *t*-test found no significant variance for any of the four *TKT* impact variables between the two groups. This suggests that something other than a dissipation in *TKT* impact over time is responsible for the variation.

**Summary**

In conclusion, some recommendations are now provided on how *TKT* could be used in, and adapted for, the Chinese context. First, the data suggested that teachers positioned lower in the educational hierarchy, as indicated by education sector and level, benefited more from participating in *TKT*. It may be the case that *TKT* is best suited for elementary and primary teachers and another programme should be used for those from the tertiary sector. Second, it would appear that the benefits to teachers are increased if the programme is completed in less than a year, so the recommendation could be made to teachers to not space out the study time between modules too greatly. Finally, the absence of a significant difference between teachers who prepared for *TKT* autonomously versus those who took part in the formalised course within China, together with the difference in impact where the formal classes were taken abroad may indicate that the effectiveness of training programmes in China could be further improved, or that the test does not require much preparation depending on who is taking it.

**Contextual factors and TKT impact**

The items in Section Three of the questionnaire were used to answer the second research question regarding the roles of contextual factors in shaping *TKT* impact. The contextual factors were those defined by Ajzen's TPB, that is, teachers' beliefs and attitude, the influence of teachers' important referents, and teachers' perceived behaviour control. Examining teachers' normative beliefs first, the response to item 37 indicated that teachers felt that they received strong encouragement from their leaders to introduce the skills and knowledge they had acquired from taking part in *TKT* into their classroom teaching, with 77% of all teachers expressing agreement. 70% of teachers also indicated that their students had responded positively to changes in classroom teaching they had introduced as a consequence of *TKT*. The level of interest from workmates, while still positive, was considerably lower with only 63% of respondents indicating that their colleagues were keen to find out more about what had been learned through *TKT*. Based on the rationale of TPB such high levels of support from referents considered important to the teachers would encourage them to introduce or implement what they had learned through *TKT*, which increased the strength of the impact of *TKT* on teaching.

Moving the discussion to teachers' beliefs about the control factors that may facilitate or impede the application of *TKT*, the data was somewhat more mixed. On the positive side, 67% of teachers indicated that what they learned through *TKT* was consistent with the National English Curriculum Standards published by the Ministry of Education which, amongst other things, specifies the type of teaching practices that teachers should follow. In a centralised education system like China such standards carry much influence. For Chinese teachers, adopting practices that are contrary to those Standards would prove problematic, so the relatively high- level agreement of the compatibility of *TKT* with the National English Curriculum Standards suggests that *TKT* practices are more likely to be adopted by teachers in China.

67% of respondents also indicated that they believed that the knowledge, skills and practices they acquired through *TKT* were applicable to their teaching context. This level of practicality however is somewhat contradicted by the relatively high levels of agreement about impediments to using what was learned in *TKT* due to the assessment system (33%) and large classes (30%). This level of concern about the examination is very understandable in the Chinese education context. Assessment across all three tiers of Chinese education is almost entirely exam based.

Correlation analysis was carried out to more clearly understand the direction and strength of any possible relationship between the respondents' teaching environment and their response to *TKT*. The tests used in this section were Pearson's r (which treated the four impact variables as scalar data) and Spearman's rho (assuming that the impact variables were ordinal) correlation analysis, which was carried out on the 10 Likert items for each of the four *TKT* impact variables. As in the previous section, the findings using both the parametric and non-parametric test were largely consistent both in terms of which associations were significant, the strength of relationships and their orientation (positive/ negative). Reporting will therefore be limited to the Pearson's r test (see Table 3). In the following discussion we use the Salkin benchmark scale[5], which is taken from Field (2013), to make judgements on the correlation metric.

[5] The scale is as follows: (+−).8 to (+−)1.0 (very strong relationship); (+−).6 to (+−).8 (strong relationship); (+−).4 to (+−).6 (moderate relationship); (+−).2 to (+−).4 (weak relationship);.0 to (+−).2 (weak or no relationship).

**Table 3: Correlation analysis**

| TPB component | Item | Subject matter knowledge | | General pedagogic knowledge | | Pedagogical content knowledge | | Context knowledge | |
|---|---|---|---|---|---|---|---|---|---|
| | | Pearson's r | Sig. | Pearson's r | Sig. | Pearson's r | Sig. | Pearson's r | Sig. |
| **Behavioural beliefs and attitudes** | 42. consistent with own teaching beliefs | 0.483 | < 0.01 | 0.675 | < 0.01 | 0.669 | < 0.01 | 0.648 | < 0.01 |
| **Normative beliefs and subjective norms** | 37. school encouragement | 0.399 | < 0.01 | 0.457 | < 0.01 | 0.463 | < 0.01 | 0.491 | < 0.01 |
| | 38. positive student response | 0.651 | < 0.01 | 0.719 | < 0.01 | 0.755 | < 0.01 | 0.703 | < 0.01 |
| | 39. colleague interest | 0.55 | < 0.01 | 0.565 | < 0.01 | 0.651 | < 0.01 | 0.617 | < 0.01 |
| **Control beliefs and perceived behavioural controls** | 40. adequate school resources | 0.41 | < 0.01 | 0.458 | < 0.01 | 0.66 | < 0.01 | 0.505 | < 0.01 |
| | 41. consistent with National English Curriculum Standards | 0.424 | < 0.01 | 0.515 | < 0.01 | 0.535 | < 0.01 | 0.467 | < 0.01 |
| | 43. *TKT* practical for my teaching context | 0.554 | < 0.01 | 0.694 | < 0.01 | 0.698 | < 0.01 | 0.716 | < 0.01 |
| | 44. language obstacle | −0.321 | < 0.01 | −0.288 | < 0.01 | −0.27 | < 0.01 | −0.26 | < 0.01 |
| | 45. assessment obstacle | −0.37 | 0.037 | −0.146 | < 0.01 | −0.188 | < 0.01 | −0.198 | < 0.01 |
| | 46. class size obstacle | −0.036 | 0.605 | −0.1 | 0.151 | −0.173 | < 0.05 | −0.222 | < 0.01 |

From the correlation analysis a more nuanced picture of the relationship between a teacher's educational context and their response to the *TKT* emerges. Looking firstly at the group of variables representing the opinions or attitudes of individuals or groups the Chinese teachers viewed as most important or influential in their educational context (students – item 38, co-workers – item 39, and leaders – item 37), a statistically significant positive correlation existed between each of the variables and the four dimensions of *TKT* impact. In terms of rank, 'Student response' (item 38) had the highest correlation with all four *TKT* impact dimensions, excepting context knowledge, where it ranked second. Across all four impact dimensions a strong positive correlation of 0.65–0.76 was found. Teachers' perception of the support they received from colleagues (item 39) was strongly correlated with pedagogical content knowledge (0.651) and context knowledge impact (0.617), with moderated positive association with the other two dimensions. School leader support had a statistically significant moderate positive correlation with the four *TKT* impact dimensions/variables.

Interpretation of this data is more problematic as correlation does not indicate the direction of the relationship between the variables, and as such cannot be used to attribute cause in the variation of *TKT* impact. All that can be definitely said is that for each of these variables a statistically significant, moderate to strong positive (and largely linear) relationship exists with the levels of *TKT* impact.

In terms of explanation of this pattern of association, it could be hypothesised that, for instance, the high levels of interest from a teacher's colleagues in what they had learned in the *TKT* course acts a stimulus to that teacher, increasing their confidence in the efficacy of the approach and making it more likely that they will put what he or she has learned into practice. This would be a hypothesis that fits in with the TPB model. An alternative explanation could be that a teacher who acquires a strong grasp of innovative teaching practices through preparation for *TKT* (as represented by pedagogical content knowledge and general pedagogic knowledge) and consequently uses them more widely and more effectively, is more likely to attract the interest of their colleagues, and is more eager to see these practices in operation and learn from them. Similar arguments hold for the other correlations described above. Receptiveness by students to alternative teaching practices, particularly ones not in accord with the existing educational culture, makes the introduction of those practices into the classroom far easier (a TPB hypothesis). An equally plausible interpretation is that a greater understanding of such practices acquired by a teacher through *TKT* makes their use in teaching more effective and interesting, thereby stimulating the interest of students.

Turning to the variables representing TPB controls on behaviour, which potentially facilitate or impede a person's ability to exhibit a behaviour desired by the change agent, the results indicate that these were far less influential in shaping *TKT* impact (or alternatively being shaped by *TKT*). The variables of class size (item 46)[6], the assessment system (item 45), and teachers' language level (item 44) had very weak negative correlation to the impact variables. Moderately positive relationships existed between the impact and teachers' views that the practices acquired through *TKT* were consistent with what was expected of Chinese teachers as outlined in the National English Curriculum Standards. If the premises of the TPB are accepted, this would appear to suggest that the factors of class size, the assessment system and teachers' language level did not play a substantial role in shaping the *TKT* impact.

[6] Not statistically significant for General pedagogic (with pearson's r but statistically significant with Spearman's rho) and Subject matter knowledge (not significant with either measure).

## Conclusions

In this study, we have investigated the scope and intensity of *TKT* impact on teachers' teaching beliefs, knowledge and practice, and the roles of contextual factors in shaping *TKT* impact through carrying out focus group interviews and a survey. The investigation of the first research question revealed that the impact was broadly positive, particularly in relation to the teachers' understanding of teaching resource use, teaching methods and skills. In terms of actual changes in behaviour, it was reported that the changes were most evident in the diversity of teaching materials used, the more considered planning of lessons and classroom activities, and the types of classroom interactions and student work modes such as individual, pair work, group work, choral work, etc. The focus group findings were supported by survey findings. The quantitative findings indicated that the test has overwhelmingly strong positive influence on the teachers' four areas of knowledge, except for subject matter knowledge. The areas most influenced by *TKT* were teaching abilities and changes in actual teaching behaviour. Specifically, *TKT* has greatly improved the teachers' understanding of teaching aims, lesson planning, teaching material evaluation, identifying student needs, use of teaching resources, teaching skills and learner assessment. In addition, there was a substantial improvement in the teachers' understanding of teaching theory, teaching methods and teaching aims. The impact on subject matter knowledge is still positive but relatively low. *TKT* has far less influence or effect on improving teachers' understanding of English language systems such as grammar.

To investigate the second research question, which aimed to draw out insights of the roles of contextual factors in shaping *TKT* impact, we firstly carried out a variation analysis to determine whether there was any significant variation in *TKT* impact amongst different respondent groups based on the demographic data, and then conducted a correlation analysis between *TKT* impact and the three contextual factors (teachers' beliefs and attitudes, the influence of teachers' important referents, and teachers' perceived behaviour control). The findings revealed that education level and sector did not have strong influence on *TKT* impact. The only statistically significant finding was for pedagogic content knowledge as there is lower impact for teachers with Master's degrees. *TKT* has stronger impact in terms of subject matter knowledge, pedagogical content knowledge and context knowledge on the teachers at the lower end of the education sector. Moreover, *TKT* has less impact for teachers who take preparation courses that are longer than one year than those who complete the course in one year or less. *TKT* has stronger impact for teachers who take part in training abroad, either entirely or partially.

With regard to correlation analysis, we concluded that the factors facilitating positive *TKT* impact included high levels of support from the teachers' referents, e.g. school leaders, students and colleagues, the consistency between the construct of the *TKT* and the National English Curriculum Standards, and the practicality of some teaching suggestions from *TKT*. Those factors which impeded teachers' application of *TKT* in practice included the assessment system and large class sizes. In addition, we found that *TKT* impact had close correlation with the teachers' referents' attitudes towards what they had learned through *TKT*, while the factors such as the assessment system, large class size, and the consistency between *TKT* and the Standards did not play a substantial role in shaping the *TKT* impact.

## Implications

### Implications for Cambridge English Language Assessment

*TKT* is generally recognised as a good test with a positive impact in China. With respect to maximising the utility of *TKT*, the findings suggest that it can be increased by targeting pre-service teachers or in-service teachers working at elementary or secondary schools. The benefit for tertiary sector teachers as well as those teachers holding higher degrees appears to be less pronounced.

### Policy-makers

The findings related to *TKT* training duration, teachers' education sector, and their education level have implications for Chinese education policy-makers. According to the findings, the *TKT* programme is more suitable for elementary or secondary school teachers. As public money is used to subsidise teachers' participation in *TKT* (as it is in most cases) the net return on that investment would be maximised (in terms of improvements in teacher quality) if teachers at the lower end of the hierarchy are targeted. *TKT* training courses abroad appear to be more effective than *TKT* preparation carried out by teachers in China, whether independently prepared or done through organised courses. The lack of difference in impact between teachers who prepared independently and those who took preparatory courses within China indicates there may be a problem with Chinese-based *TKT* training programmes. If the *TKT* is to be more widely used in China then the standards of domestic organised training need to be improved. Where training is organised, either domestically or abroad, the preparation should be relatively intensive and completed within a year for maximum benefit.

### Teachers

The present study revealed that *TKT* has broadly positive impact on teachers' teaching beliefs, knowledge, behaviour and their own language development. We can conclude that pursuing a *TKT* certificate is one of the effective ways to improve teachers' academic and practical knowledge and skills.

## Limitations

There are some limitations to this research. Firstly, teachers sitting *TKT* in China are typically sponsored by their schools or districts and those selected to take part are based on success in a competitive exam. As such, they are often drawn from a younger, more highly educated cohort of teachers. We might claim that the teachers taking part in *TKT* (i.e. the population from which our sample comes) is not typical of the general population of English language teachers in China. Therefore, future research could focus on people who study or sit *TKT* independently.

Secondly, while correlation analysis cannot be used to evaluate causation, it still provides an understanding of the strength and orientation of relationships. It therefore provides some insight into the utility of TPB as a model for stakeholder response to a test. The analysis of TPB variables we described indicates that the TPB variables function in the direction anticipated by the model. Future research, using regression analysis of better operationalised indicators of the three dimensions of TPB belief, could be used to develop a model of test impact on stakeholders such as teachers.

Although the phenomenon 'impact of a test' has been widely investigated within different contexts by researchers, most of the studies focus on examinations for students. The impact of an examination for teachers is explored far less rigorously. To the best of our knowledge, the present study is the first one which involved such a large number of *TKT* test takers across China. This paper makes no claim to having formulated the definitive approach either in conceptualising *TKT* impact (the four aspects of teaching knowledge *TKT* covers), nor in its operationalisation (the variables selected to represent those aspects). But it does provide an initial attempt at devising a tool that, once refined, will allow research into the impact of *TKT* on teaching to be more generalisable and verifiable, characteristics of quantitative research. Further research into other contexts to explore analytical frameworks and analysis methods would be a further positive benefit of this research.

## References

Ajzen, I (1985) From intentions to actions: A theory of planned behavior, in Kuhl, J and Beckman, J (Eds), *Action-control: From Cognition to Behavior*, Heidelberg: Springer, 11–39.

Ajzen, I (1991) The theory of planned behavior, *Organizational Behavior and Human Decision Processes* 50, 179–211.

Ajzen, I (2006) *TPB Diagram*, available online: people.umass.edu/aizen/tpb.diag.html

Cheng, L Y (2005) *Changing Language Teaching through Language Testing: A Washback Study*, Studies in Language Testing volume 21, Cambridge: UCLES/Cambridge University Press.

Creswell, J W and Plano Clark, V L (2011) *Designing and Conducting Mixed Methods Research*, Thousand Oaks: Sage.

Field, A P (2013) *Discovering Statistics using IBM SPSS Statistic: And Sex and Drugs and Rock 'n' Roll*, 4th edition, London: Sage.

Huang, L Y (2011) *Washback on Teachers' Beliefs and Behaviour: Investigating the Process*, Beijing: Foreign Language Teaching and Research Press.

Ministry of Education of the People's Republic of China, (2001) 国家英语课程标准 [National English Curriculum Standards], Beijing: Beijing Normal University Press.

Spratt M (2004) *The Construct of TKT*, Cambridge ESOL Report No. 0634.

Valazza, G (2008) Impact of TKT on language teachers and schools in Uruguay, *Research Notes* 4, 21–26.

Wall, D (1997) Test impact and washback, in Clapham, C and Corson, D (Eds) *Encyclopedia of Language and Education Volume 7: Language Testing and Assessment*, Dordrecht: Kluwer Academic Publishers, 291–302.

Wall, D (2005) *The Impact of High-stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*, Studies in Language Testing volume 22, Cambridge: UCLES/Cambridge University Press.

## Appendix 1

TKT Impact Questionnaire Survey

Dear Teachers,
Thank you for your participation in this survey. This questionnaire survey is designed to study how the *TKT* influences your teaching beliefs, knowledge and practice. Information is gathered for our research project only. Kindly please complete the survey according to the actual situation.
Thanks again for your sincere cooperation and assistance!
Yours sincerely,
Liyan Huang and Angelo Papakosmas

### Section One: Demographic data

(From item 1 to item 8. Please choose the answer that best describes your situation.)

1. Age
   a. 20–30 b. 31–40 c. 41–50 d. 51+
2. Teaching experience
   a. no more than 5 years b. 5–10 years c. 10–20 years d. more than 20 years
3. Education level
   a. High school degree b. Secondary vocational school education c. Post-secondary education d. College degree e. Master's degree f. Doctoral degree
4. Education sector
   a. Public University b. Public High School c. Public Junior High School d. Public Elementary School e. Private University f. Private High School g. Private Junior High School h. Private Elementary School
5. Region of China
   a. Beijing b. Guangdong c. Hebei d. Xinjiang e. Others
6. Way of preparing for *TKT*
   a. To participate in training in China b. To participate in training abroad c. To participate in training at home and abroad d. Self-study
7. *TKT* preparation duration
   a. One month b. Two months c. Three months d. Six months e. A year f. Other
8. Time for sitting *TKT*
   a. 2010 b. 2011 c. 2012 d. 2013 e. before 2010

### Section Two: TKT impact

(From item 9 to item 36. Please choose the answer that best describes your personal opinions.)

a. Strongly agree b. Agree c. Neither agree nor disagree d. Disagree e. Strongly disagree

After completing the TKT preparation and sitting TKT, I . . .

9. have a better understanding of English grammar.
10. have a better understanding of English vocabulary.

11. have a better understanding of English phonetics.
12. have a better understanding of listening skills and sub-skills.
13. have better understanding of speaking skills and sub-skills.
14. have better understanding of reading skills and sub-skills.
15. have better understanding of writing skills and sub-skills.
16. have better understanding of different teaching methods.
17. have better understanding of teaching theory.
18. have a greater awareness of the positive aspects of my teaching.
19. have a greater awareness of the negative aspects of my teaching.
20. am better able to evaluate teaching materials.
21. am better able to plan for my student needs.
22. am able to better plan my lessons.
23. use a greater variety of classroom activities in my teaching.
24. My students enjoy my lessons more.
25. have a better understanding of different learner styles.
26. have improved language knowledge teaching skills.
27. have more ideas to develop my students' listening strategies.
28. have more ideas to develop my students' speaking strategies.
29. have more ideas to develop my students' reading strategies.
30. have more ideas to develop my students' writing strategies.
31. have better understanding of vocabulary related teaching methods.
32. have better understanding of grammar related teaching methods.
33. am more able to choose appropriate assessment activities for learners.
34. have better understanding of teaching aims.
35. can identify my students' needs more precisely.
36. am able to use the teaching resource in our school more effectively.

Section Three: Contextual factors

(From item 37 to item 46. Please choose the answer that best describes your personal opinions.)
a. Strongly agree b. Agree c. Neither agree nor disagree
d. Disagree e. Strongly disagree

37. My school leaders have encouraged me to put what I have learned through *TKT* into practice.
38. Students have responded positively to the changes I have introduced.
39. Fellow teachers have expressed an interest in finding out more about what I have learned through *TKT*.
40. My school is equipped with the resources to allow me to put what I have learned into practice.
41. What I learned through *TKT* is consistent with the aims of the National Standards.
42. What I learned through *TKT* was consistent with my own ideas about teaching.
43. What I learned through *TKT* is practical for teaching in my situation.
44. My English language level restricts my ability to implement what I have learned through *TKT*.
45. Our exam system hinders the use of *TKT* teaching ideas.
46. Large class size has restricted my ability to implement what I learned from *TKT*.

# Investigating the face validity of Cambridge English: First in the Cypriot context

**DINA TSAGARI** DEPARTMENT OF ENGLISH STUDIES, UNIVERSITY OF CYPRUS

## Introduction

Various professional organisations stress the importance of monitoring the impact generated by high-stakes language tests. For instance, the European Association for Language Testing and Assessment (EALTA) Guidelines for Good Practice in Language Testing and Assessment (www.ealta.eu.org/documents/archive/guidelines/English.pdf), the International Language Testing Association (ILTA) Code of Ethics (www.iltaonline.com/code.pdf) and the Draft Code of Practice (www.iltaonline.com/ILTA-COP-ver3–21Jun2006.pdf) advise test designers to investigate the consequential validity of their tests including intended and unintended washback effects on teaching and learning (Alderson and Wall 1993, Bailey 1996) as well as the ethical and social effects their tests have on the wider community (Canale 1987, Hughes 2003, McNamara and Roever 2006). However, while aspects of consequential validity such as 'washback' (that is 'the influence of testing on teaching and learning', see Alderson and Wall 1993) have received attention in recent years (e.g. Cheng 2005, Hawkey 2006, Tsagari 2009, Wall 2005), stakeholders' judgements about tests (referred to as 'face validity'), especially those operating in educational contexts for long periods of time, have not yet been systematically investigated in the field.

Face validity is generally defined as the appeal of a test or the judgements made about the appropriateness of a test by potential test takers and test users (Alderson, Clapham and Wall 1995, Bachman and Palmer 1996, Hughes 2003). Face validity of tests is considered important in language testing and teaching as it can influence the attitudes of teachers and students (among other stakeholders) towards tests and the adaptation or not of the tests (Hughes 1989:27).

Alderson et al (1995) also maintain that test takers, when confronted by lack of face validity of a test, may not perform as well as they could. Investigating face validity adds to the test's validity, especially consequential validity (Alderson and Wall 1993, Bachman and Palmer 1996, Bailey 1996, Hughes 1994, Messick 1996) and reflects an examination board's commitment to professional, ethical, and legal accountability towards its stakeholders, such as test takers, teachers, employers, university admission officers, etc. (Hamp-Lyons 1997, Rea-Dickins 1997, Taylor 2000). Finally, Rea-Dickins (1997) stresses that ensuring face validity puts stakeholders at the heart of assessment, democratises assessment processes and promotes greater fairness. It also makes the examination more appropriate for particular candidates, contexts and test uses.

Stakeholders such as teachers, in particular, have unique insight into the appeal and effects of tests. They are in a position to advise students which tests to take, prepare them for these tests, and can see how testing affects students in their day-to-day lives. Therefore, teachers are well positioned to recognise discrepancies between classroom and test practices and can gauge the effects of the tests chosen on students (Norris 2008).

Against this background the present study aimed to explore the face validity of *Cambridge English: First*, also known as *First Certificate in English* (FCE). The study is based on teachers' feedback on the aforementioned exam in the educational context of the Republic of Cyprus (hereafter referred to as Cyprus) where the exam has been operating for more than a decade.

## Description of context and research questions

English has a prominent place in Cyprus not only because of its colonial past but because of the significant role that English has acquired as a current *lingua franca*. Cyprus has developed a long tradition in the teaching and learning of English, which was introduced officially in the state educational system in 1935 and has been taught almost continuously since then with some short interruptions due to political circumstances. English as a Foreign Language (EFL) is nowadays taught at the pre-primary (on a pilot basis), primary, secondary and tertiary level of education in state-funded schools and institutions of higher education. However, no official standardised testing or certification system is used in the state sector to accredit the level of EFL students. English is assessed on the basis of teacher-made progress/achievement tests or the National Exam for university entry, the latter being controlled by the Ministry of Education and Culture (MoEC) of the country. In addition, there are a few private schools which use English as the medium of instruction across all school subjects taught (English as a Second Language – ESL). The popularity of these schools is currently on the increase. At the same time, English is taught and learned at the hundreds of private foreign language schools (also known as *frontistiria*) in the country and there is also a considerable amount of private teaching of the language, too, e.g. one-to-one tuition at a student's or teacher's home (for further discussion see Karoulla-Vrikki 2013, Lamprianou 2012, Lamprianou and Afantiti Lamprianou 2013, Papadima-Sophocleous 2013, Pavlou 2010).

Various external standardised English language exams have been operating in the private EFL sector, including *Cambridge English: First* (www.teachers.cambridgeesol.org/ts/digitalAssets/117578_Cambridge_ English_First__FCE__Handbook.pdf) which has been recognised as a measure of language proficiency by various educational and professional organisations in the country. However, even though many hundreds of Cypriot students take the exam every year, we know little about local stakeholders' perceptions of the exam, its use in the local context and its impact on teaching and learning. For example, discussions with teachers showed that the exam is considered demanding in terms of its Use of English and Speaking component (see Tsagari 2012a). Also official exam results (2004–12) show that the overall pass rate of Cypriot candidates is indeed low, e.g. less than 50% acquire *Cambridge English: First*, usually with a 'C' grade (see Table 1).

Therefore the aim of this study was to investigate the face validity of *Cambridge English: First* to arrive at a clearer understanding of teachers' perceptions of the utility, efficiency, practicality, difficulty, content validity and washback of the exam within the context of private language education in Cyprus. The research question the present study addressed was the following: 'What are the perceptions of teachers towards *Cambridge English: First?*'

**Table 1: Exam results of Cambridge English: First for Cyprus (2004–12)***

| Grade A | Grade B | Grade C | Total pass | Grade D | Grade E |
|---|---|---|---|---|---|
| 4.37% | 8.36% | 32.3% | 49.52% | 14.71% | 40.38% |

* www.cambridgeenglish.org/research-and-validation/quality-and-accountability/grade-statistics/

## Research design

### Method

To answer the research question, open-ended in-depth interviews were conducted with both individual EFL teachers preparing students for the exam and focus groups (Bloor 2001, Krueger and Casey 2009, Litosseliti 2003). The interviews were based on an interview guide (specially designed for the purposes of the study and informed by the relevant literature) that covered a range of aspects of the face validity of the exam (see Appendix 1). The interviews were recorded with the agreement of the participants and were carried out between November 2011 and April 2012. The study also analysed characteristics of the educational context and those of the teachers that prepare students for *Cambridge English: First*.

### Participants

The interviews took place in 10 schools. Nine of them were *frontistiria* and one was a private English-medium school. The teachers who participated in the study were all well experienced in preparing students for various high-stakes English language exams and had a Bachelor of Arts (BA) degree in English (Linguistics or Language and Literature) as their minimum teaching qualification. All teachers said they were very familiar with *Cambridge English: First* (they had 6–10 years of experience preparing students for the exam) and other Cambridge English exams (see Table 2). Students preparing for *Cambridge English: First* were between 14 and 16 years old and received on average five years of English language tuition prior to taking the exam.

**Data collection and analysis**

The approach to data gathering and analysis was qualitative. The interviews were transcribed verbatim to provide an accurate record of spoken language. For the analysis of the data a special statistical package was used, e.g. ATLAS.ti 6. Teachers' responses were inspected for common or divergent themes and analysed through an inductive approach during which themes and patterns emerged from the data (Paltridge and Phakiti (Eds) 2010). The analysis identified several factors influencing participants' perceptions of the exam. These are presented in Table 3 as subthemes that emerged from the qualitative interview data.

**Table 2: Teachers' profile**

| Teachers | Gender | Town | Cambridge English: First | PET* |
|---|---|---|---|---|
| Teacher 1 | Male | Limassol | ✓ | ✓ |
| Teacher 2 | Female | Paralimni | ✓ | – |
| Teacher 3 | Female | Larnaca | ✓ | – |
| Teacher 4 | Female | Larnaca | ✓ | – |
| Teacher 5 | Female | Limassol | ✓ | – |
| Teacher 6 | Female | Limassol | ✓ | ✓ |
| Teacher 7 | Female | Limassol | ✓ | – |
| Teacher 8 | Female | Nicosia | ✓ | – |
| Teacher 9 | Female | Nicosia | ✓ | ✓ |
| Teacher 10 | Female | Nicosia | ✓ | – |
| Teacher 11 | Female | Nicosia | ✓ | – |
| Teacher 12 | Female | Limassol | ✓ | ✓ |
| Teacher 13 | Female | Nicosia | ✓ | – |
| Teacher 14 | Female | Larnaca | ✓ | – |
| Teacher 15 | Female | Nicosia | ✓ | – |

** Preliminary English Test (now known as Cambridge English: Preliminary)*

**Table 3: Coding categories and themes**

| Major category | Subtheme |
|---|---|
| 1. Overall appreciation of the examination | • Positive comments |
| | • Negative comments |
| | • Logistics: Time limits |
| 2. Usefulness of the examination | • Difficulty/level |
| | • Status/recognition |
| | • Cost |
| 3. Exam preparation materials | • Use |
| | • Availability |
| | • Features |
| 4. Impact of the exam on preparation (or learning and teaching?) | • Instruction |
| | • Curriculum |
| | • Learning |
| | • Accountability |

A sociocultural theory perspective was employed to interpret the data. Sociocultural theory has recently been having a significant impact on the interpretation of classroom experiences and practices and on the analysis of the development of language skills (Kramsch (Ed) 2002, Lantolf (Ed) 2000, Lantolf and Thorne 2006). Informed by this theoretical approach, the findings of the present study are interpreted and reflected upon through the realities of the local society and culture these occurred in. Factors influencing teachers' practices as well as their beliefs of the importance of EFL, the role of private education in Cyprus and the importance of language certificates in the Cypriot society were taken into consideration when the data was analysed.

The next section presents the results of this study. It is organised according to the major themes arising from the data analysis and key points are exemplified through teachers' interview quotes.

## Results

### Overall appreciation of Cambridge English: First

*Positive comments*

The majority of the teachers (N = 10) were positive about *Cambridge English: First*. Teachers generally perceived *Cambridge English: First* as being a good exam that provides an opportunity to effectively develop students' language skills because the exam has good skills coverage, e.g.:

> Teacher 12: *Well, we stick to Cambridge exams. Yes, we support Cambridge exams and we feel that they prepare candidates really well. Cambridge English: First gives them a lot! Students benefit from the exam preparation in terms of knowledge they get . . . real knowledge of the English language . . . Absolutely, yes!*

Other than skills coverage, teachers (N = 6) also stressed that the exam constitutes a challenge for both students and teachers, e.g.:

> Teacher 9: *The exam itself has got everything . . . from grammar to listening and speaking. It is very demanding but at the same time very effective . . . It's very challenging even for us . . . I mean I need to come up with different lessons each time.*

Other teachers (N = 4) spoke highly of specific skills such as writing, e.g.:

> Teacher 12: *I definitely like the fact that they practise different kinds of writing in the Cambridge English: First. So students get to know how to work on an article, on an essay, etc.*

*Negative comments*

Some teachers (N = 3) commented on the lack of integration of skills in *Cambridge English: First*, e.g.:

Teacher 2: *I consider Cambridge English: First just teaching bits without any coherence and in isolation.*

Other teachers (N = 4) expressed doubts with regard to the emphasis placed on certain aspects of language, e.g.:

Teacher 14: *In Cambridge English: First, I think, there is a lot of grammar. Students are not so much into grammar. They find it very difficult to remember all these rules and things they have to do.*

Two of the teachers also expressed doubts concerning the validity of students' results. For example, Teacher 14 felt frustrated when her predictions on how well her students were going to do in the exam were often counter-intuitive, e.g.:

Teacher 14: *I have had cases of good students who failed the exam. Sometimes it is the difficulty of the readings or maybe the way they mark the exam . . . I cannot understand why a student I didn't believe could make it actually passed the exam. And the student that I was so confident about failed the exam. There should be an explanation.*

*Logistics: Time limits*

Two of the teachers expressed concerns about the logistics of the exam. Teacher 14 said: *It is quite a long exam. Students are sometimes bored with it*. Teacher 13 also noted that she does not approve of the strict time limit students have to complete the exam papers.

Teacher 13: *The thing I do not like about this exam is the time limit . . . I think it is just nerve-racking for the students. I believe that 15 extra minutes per paper would have been better.*

**Usefulness of the exam**

*Difficulty/level*

The low pass rate of Cypriot candidates (see also Table 1) seems to have become a preventive factor for certain teachers and students, e.g.:

Teacher 2: *We used to do Cambridge English: First but we abandoned it because we found that the pass rate is low 'cause the exam has got a lot of grammar in isolation which presents difficulties to kids.*

One of the teachers believed that *Cambridge English: First* requires extra time to prepare, e.g.:

Teacher 5: *Cambridge English: First is a constant struggle . . . and . . . it takes a lot of time to prepare.*

For another teacher preparing average students for the exam accentuates the problem, e.g.:

Teacher 12: *. . . the Cambridge English: First exam is really demanding, so it requires a lot of studying on their behalf, irrespective of what I do in class. So they are not willing to go through that, especially if they are average students . . .*

One of the teachers felt that certain aspects of *Cambridge English: First* are particularly difficult for their Cypriot students, e.g.:

Teacher 1: *For Cambridge English: First there are a lot of specific things . . . like transformation sentences and certain vocabulary exercises and so on and students need to study hard for these exercises which I know they don't do.*

*Status and recognition*

Factors such as recognition and the status of the exam in the local context seem to affect teachers' attitudes towards the examination. For example, some of the Cypriot teachers (N = 7) were under the belief that Cambridge International General Certificate of Secondary Education – English as a Second Language (IGCSE – ESL) is widely recognised, unlike *Cambridge English: First*, especially when students apply for a university place abroad.

Teacher 2: *A lot of universities abroad do not accept Cambridge English: First whereas they accept Cambridge IGCSE – ESL.*

Others (N = 4) believed so because of the limited recognition of the exam by the local professional market and educational authorities.

*Cost*

The cost of exam preparation and exam fees is another factor that appears to have negatively impacted on the popularity of *Cambridge English: First*. As Teacher 13 commented:

Teacher 13: *I stopped preparing students for Cambridge English: First because I thought it was very expensive and it was not recognised. It was recognised for its value only. It has quite a lot of material that is good to use but it was too expensive for parents to pay for an exam that didn't have wide recognition . . .*

**Exam preparation materials**

*Use of Cambridge English: First preparation materials*

To facilitate exam preparation, teachers use practice tests and past papers with varying intensity, e.g.:

Teacher 1: *I start working with practice papers and past papers before Christmas . . . then I work more intensely, let's say, after January.*

Teacher 4: *I start working with practice tests for the Cambridge English: First as early as October. Like a year before I give them the first one. Yes, I start in October giving them the first one so that they get a taste of what it means . . . you know . . . they get a very good idea especially when they get a low mark . . . they*

> *know they have to work hard . . . much harder . . . and then, after a month they get a second one . . . it gradually becomes more intense throughout the year . . . yes. . . .*

> Teacher 6: *We start teaching from past papers from year 4 to show them a few papers during that year and guide them through the papers. From the beginning of year 5 we work with our textbooks but every now and then we'll give them past papers and we will give them past papers as exams and as tests, and then, in January, before they take the exam, we work solely from past papers.*

*Use of Cambridge English: First in the service of other exams*

Unlike the perceived difficulty and recognition of *Cambridge English: First*, the actual preparation for the exam as well as the teaching materials used were highly valued. Teachers believed that these constitute the key to effective English language learning and success in other Cambridge language exams. For example, in the case of Teacher 14, it was highly pertinent for students to prepare for *Cambridge English: First* a year before they prepare for the Cambridge IGCSE – ESL.

> Teacher 14: *I think if a student has taken Cambridge English: First classes he can understand more easily and adjust to the Cambridge IGCSE – ESL level. But if you have a student who hasn't prepared for Cambridge English: First he finds Cambridge IGCSE– ESL more difficult. So I would say that Cambridge English: First is the step to the Cambridge IGCSE – ESL. I would call this the Pre-IGCSE level. Yes! . . . From my experience I find that students who prepare for Cambridge English: First the year before, do much better in the Cambridge IGCSE – ESL.*

Some other teachers would use *Cambridge English: First* materials when preparing for other Cambridge exams. For example, Teacher 1 uses *Cambridge English: First* materials to prepare students for *Cambridge English: Preliminary* (also known as *Preliminary English Test (PET)*).

> Teacher 1: *I use the Cambridge English: First listening and speaking book at the PET level . . . I try to find books that are not exactly PET, a bit, you know, a bit more advanced.*

Two teachers said they use materials from the Use of English paper when they prepare students for the local University Entrance Exam, e.g.:

> Teacher 14: *I also use some Cambridge English: First materials to prepare students for the Pancyprian Exam [University Entrance Exam] as well. Because they are very helpful, especially transformations . . .*

*Availability of preparation materials*

What seems to have made teachers resort to the use of *Cambridge English: First* materials is the limited production and lack of quality of preparation material for other exams, e.g.:

> Teacher 3: *. . . that's a big problem. There isn't enough material available for Cambridge IGCSE – ESL. Unfortunately the publishers aren't interested in publishing materials for Cambridge IGCSE – ESL because Cyprus is not a big market. For example, Greece is a larger market so publishers produce a lot of materials for Cambridge English: First for the Greek market and they are not really interested in the small Cypriot market. Basically, we've got only a handful of books for Cambridge IGCSE – ESL and these are not really satisfactory.*

**The impact of Cambridge English: First on teaching and learning**

*Instruction*

Teachers talked openly about their exam preparation practices. Teachers admitted that they 'teach to the test' when preparing students for *Cambridge English: First*, e.g.:

> Teacher 2: *. . . in the Cambridge English: First year we just do exam practice.*

> Teacher 3: *We do not deviate . . . we do not do anything else apart from . . . the content of the exam. The Cambridge English: First exam influences my teaching to a great extent.*

Teacher 4, in particular, teaches each skill in isolation when preparing for the exam. The teacher explained that she does so because that is the way to succeed in the exam, e.g.:

> Teacher 4: *We aim for a good grade in the exam . . . so I do each part separately, only reading, only listening, only writing and so on . . . to teach them what I want, what my aims are and then I focus on the past papers.*

Other teachers would go as far as to encourage students to learn language by heart, drawing heavily on past exam papers, e.g.:

> Teacher 1: *I use a lot of past papers and I prepare some handouts, presenting for example the expressions they need when they write a summary, the expressions they should use when they write an informal letter or a formal letter. I ask them to memorise those things, learn them by heart and apply them whenever it's possible. The same with speaking. That's exactly what we do with speaking. I give them certain expressions, for example how to move on from one point to the other or when they give the speaking card they are supposed to speak for 7 minutes or so. Then, I give them different expressions to use when they want to give an example, you know . . . 'for example', 'for instance', . . . so I basically do this with handouts.*

*Learning*

Four of the teachers were aware that their teaching is restricted to exam requirements and that is likely to produce negative washback on learning, e.g.:

Teacher 1: *My teaching is very dependent on the exam. What I teach is very much related to the exam. This is not very good of course. Students come here to learn English and that's the aim for them: to learn English, to communicate in various situations and so on.*

One of the teachers thought that exam results are not indicative of students' language knowledge and abilities, but rather of their exam preparation skills and strategies, e.g.:

Teacher 12: *I don't think it's a matter of knowledge in exam preparation. It's just a matter of skills, strategies, nothing else! For example, as far as the Cambridge English: First is concerned, I keep reminding them of how to work in each part of the exam. That you have to do this . . . and this . . . and this . . . Yes. And then I ask them: What do we do when we have a passage in front of us? I mean we don't read the whole thing. We skim and scan. All the time! Yes because I think that's important! That's what the Cambridge English: First wants! It doesn't test knowledge of the language.*

Teacher 5 noted that her exam-oriented approach is justified in the face of exam success, e.g.:

Teacher 5: *We try to avoid that but unfortunately there are days where yes, the lesson has to be boring going through past papers and exam tips . . .*

Teacher 13 also acknowledges that she has to employ an exam-oriented approach but she believes that this is not a problem or threat to the quality of learning, e.g.:

Teacher 13: *OK, I believe they benefit from the preparation . . . to pass the exam, OK, definitely! Definitely! Depends on the preparation you're doing of course! But, preparing a student for an exam, 100% sure they benefit from that . . . OK. Without realising it of course, you leave out important things . . . especially towards the exam . . . you leave out other things that you would have liked to do in the class . . . because you are preparing your students for that particular exam. But of course the Cambridge English: First has got the composition, and the listening, the oral part and the comprehension part, well you are obliged to do all these things to prepare them for the exam. So you don't leave too many things out, I believe! That's why!*

*Accountability*

Teachers want to make sure that their students do well in the exam because competition is fierce. However, teachers (and schools) are held accountable for test results. The higher the success rates of their students, the better teachers or schools they are perceived to be, e.g.:

Teacher 1: *. . . But if you have a private school like mine, you need to have results. Very good results. Therefore, you are forced to teach according to what they are expected to know in the exams. Do you understand what I mean? The aim is to pass the exam . . . I mean, they do learn English at the end of the day but yes, what I want them to do is to pass the exams with high grades. I know what you think about this, but still, this is not just a school, it's also a business.*

Even though teaching to the test seems to be a common practice among the teachers, Teacher 15, concerned about exam preparation, suggests that an overall exam preparation approach should not be test driven, e.g.:

Teacher 15: *Let me tell you something. It always depends on the teacher. That is if the ultimate goal is Cambridge English: First, then the teacher falls in the trap of having to prepare students for this exam. There should be an overall exam preparation approach.*

## Discussion of results

The goal of the study was to record teachers' perceptions towards *Cambridge English: First* and to see if those perceptions could meaningfully contribute to an understanding of the exam's face validity.

Data was gathered by interviewing teachers who had prepared students for the exam. Teachers' responses to the open-ended questions provided valuable insight into the face validity of the exam.

Overall, teachers thought highly of the exam. They believed that *Cambridge English: First* has content validity in terms of language skills and topics, and positive impact on language learning. However, as illustrated in the previous section, teachers also highlighted certain aspects of the exam that are in need of improvement. Teachers would like the exam to develop life-long learning skills, provide students with authentic opportunities for language use in everyday situations, give students a sense of achievement within appropriate time limits and effectively prepare them for university study. Nevertheless, there are certain prevailing attitudes which, even though shared by teachers, are not in line with the test providers, e.g. issues of exam status and recognition. More specifically, teachers are under the false impression that *Cambridge English: First* is not widely recognised as a language qualification and is therefore lacking in appeal for entry to international and local universities and the professional world. These misguided perceptions make *Cambridge English: First* less attractive than other language exams which have been operating in the Cypriot context, e.g. IGCSE. However, these seem to be the perceptions of teachers in Cyprus as recognition of exams such as the IGCSE are less well known or relevant in other contents. Actually *Cambridge English: First* is indeed recognised by many local and international corporations as well as a number of universities (see www.cambridgeenglish.org/recognition/results.aspx?country=United%20

Kingdom&type=All&exam=FCE#) but Cypriot teachers do not seem to be aware of this.

Another issue that seems to be playing a discouraging role in registering students for *Cambridge English: First* is the perceived high cost of the exam preparation and fees, and the level of difficulty of the exam which, as teachers claim, sound unappealing to parents and therefore prevents them from registering students for the exam. It is the case that the pass rate of Cypriot candidates is indeed low compared to the overall pass rate (see www.cambridgeenglish.org/research-and-validation/quality-and-accountability/grade-statistics/). One explanation for this could be that the Cypriot candidates sit the exam before they have reached the appropriate proficiency level or they might be too young when they take the exam. However, further investigation is needed with regard to the difficulty of *Cambridge English: First* and the low pass rate recorded for the Cypriot cohort of examinees (see Table 1).

Nevertheless, regardless of its perceived difficulty, *Cambridge English: First* seems to be valuable to teachers when preparing students for other exams. As teachers noted, *Cambridge English: First* offers students both a comprehensive revision of the previous year's and additional significant language specifics that, to their understanding, adequately prepare students for any exam at the level of *Cambridge English: First* and beyond. Therefore, teachers resort to *Cambridge English: First* materials to prepare students for other exams. It is very often the case that a whole year is devoted to the use of *Cambridge English: First* materials prior to taking other exams. What seems also to make teachers resort to the use of *Cambridge English: First* materials is the limited production and lack of quality of preparation materials for other exams, e.g. IGCSE.

With regard to the impact of the exam on teaching and learning, the findings indicate that the instructional practices of the interviewed Cypriot teachers are largely driven by the content of *Cambridge English: First*. Despite their awareness of exam impact on language learning and teaching, teachers say they use various exam-oriented techniques to meet the requirements of the exam in their effort to successfully prepare their students for it. For example, teachers said they use practice tests and past papers (at times quite early in the preparation cycle), give advice and exam tips to students as well as test-taking techniques and recommend various approaches (dubious at times, e.g. rote learning of language) that they feel will help their students do well in the exam. As teachers explained, this is due to accountability reasons and fierce competition on the market of private institutions in the country (the higher the success in the exam, the better a school or teacher is). The findings of the study also point to the presence of positive washback. This is evident in the amount of work done on all language skills including listening and speaking plus grammar and vocabulary. It seems that the exam supports a balanced approach to teaching as it encourages the development of all four skills.

## Concluding remarks

The study reported in this article explored the concept of face validity towards *Cambridge English: First*. The study also brought to light several unintended side effects of the exam and has shown how these depend on the idiosyncratic character of the local educational context. For example, the face validity of the exam seems to have been influenced by teachers' misunderstanding of the exam's underlying principles, use and status within the local private language school context. As evidenced in the data analysed, teachers also used a rather limited range of exam-preparation techniques and were unaware of teacher support for exam preparation provided on the Cambridge English website (*Cambridge English: First*: www.cambridgeenglish.org/cambridge-english-for/teachers/information-and-resources/). Finally the unavailability and low quality of preparation materials for other exams seems to be another reason for choosing *Cambridge English: First* rather than IGCSE.

Given this state of affairs, it seems reasonable to suggest that Cambridge English Language Assessment needs to establish effective channels of communication with teachers and local markets overall, e.g. disseminate further the teacher support and exam information currently available online. This will direct local teachers and other stakeholders to it and will eventually clarify misconceptions and misunderstandings of the underlying principles, use and recognition status of *Cambridge English: First* in the Cypriot context. Another way of handling the situation would be if Cambridge English could pay visits to *frontistiria* which stopped registering students for the exam, organise seminars and presentations to clarify misconceptions by developing teachers' assessment literacy and assist them in understanding their choice of exams and principles of good practice in assessment. This could be done by setting up regular meetings where teachers and school owners can ask questions, comment or contemplate on issues of interest, etc. Such initiatives can also become a springboard for the wider dissemination of Cambridge English *for Schools* examinations (developed in 2010) introduced in a small number of selected schools in the country in 2011. Test providers also need to make sure that the use of exam resources and teaching materials in the Cypriot educational context is adequate and efficient (Tsagari 2009, 2012b) and should aim positive washback and effectively disseminate their online support initiatives. Finally Cambridge English needs to consider the cost of exam fees, especially in countries such as Cyprus (and Greece) where the financial crisis is going to persist for many years to come.

The results of the present study might be relevant to other educational contexts, too. However, further studies are needed to confirm or disconfirm the present results. Nevertheless, the study has shown that seeking feedback from teachers about their test experiences provides critical information about the validity of a test. Therefore, teachers and other stakeholders (e.g. students, parents, employers, etc.) should be given a 'voice' in the assessment process. This is worthwhile because their responses provide valuable information about the pedagogical, ethical and social consequences of a test, information that might not otherwise be available. By taking stakeholders' perceptions into consideration, test providers can improve their language tests and better ensure they are valid, produce reliable scores and have positive impacts.

### Limitations and directions for further research

The study described here had two main limitations. First of all, the responders were self-selected and so were not a random sample. It is likely that teachers with strong opinions responded to the call for participation in the study. Future studies on face validity of tests should include a reasonably large sample size and use a mixed methods design to gather both qualitative and quantitative data (e.g. survey/questionnaires). Second, the study group was limited to teachers. The face validity of *Cambridge English: First* could be more thoroughly evaluated by triangulating statistical and qualitative data from teachers with data from other stakeholders such as students, parents, employers, university officers, etc. At the same time, it would be helpful to conduct similar studies in other countries where the exams are also used. Also, it might be best to have an external evaluator who would construct and conduct such studies, as has been suggested by other researchers (Winke 2011). Doing so would help the community and test providers gain a more comprehensive view of stakeholders' important contributions to the validity arguments of their tests and increase trust in the language testing products offered.

Further research is also required to address some of the teachers' comments. For example, further investigation into the factors that might have created the impression that *Cambridge English: First* is equivalent to other exams operating on the market is needed, e.g. through surveying the language entry requirements used by local professional and educational organisations, interviewing admissions officers to find out their perceptions and practices of the accreditation of language credentials, etc. In addition, comparability studies between *Cambridge English: First* and other exams offered in the local market might be another future research direction that can provide empirical evidence for the relation (if any) between these exams. Further studies could be conducted to verify the availability and investigate the quality of exam preparation materials in the Cypriot exam market as well as to clarify whether the instructional practices preferred by teachers during exam preparation are due to exam requirements or personal preference and practice.

## References

Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14 (2), 115–129.

Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

Bachman, L F and Palmer, A (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford: Oxford University Press.

Bailey, K M (1996) Working for washback: A review of the washback concept in language testing, *Language Testing* 13 (3), 257–279.

Bloor, M (2001) *Focus Groups in Social Research*, London: Sage.

Canale, M (1987) The measurement of communicative competence, *Annual Review of Applied Linguistics* 8, 67–84.

Cheng, L Y (2005) *Changing Language Teaching through Language Testing: A Washback Study*, Studies in Language Testing volume 21, Cambridge: UCLES/Cambridge University Press.

Hamp-Lyons, L (1997) Washback, impact and validity: ethical concerns, *Language Testing* 14 (3), 295–303.

Hawkey, R (2006) *Impact Theory and Practice: Studies of the IELTS Test and Progetto Lingue 2000*, Studies in Language Testing volume 24, Cambridge: UCLES/Cambridge University Press.

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

Hughes, A (1994) *Backwash and TOEFL 2000*, paper commissioned by the Educational Testing Service (ETS), University of Reading.

Hughes, A (2003) *Testing for Language Teachers*, 2nd edition, Cambridge: Cambridge University Press.

Karoulla-Vrikki, D (2013) *English-medium instruction in higher education in Cyprus: Changing politics, ideologies and language policies*, paper presented at the Nitobe Symposium Languages and Internationalization in Higher Education: Ideologies, Practices, Alternatives, Reykjavik, Iceland, 18–20 July 2013.

Kramsch, C J (Ed) (2002) *Language Acquisition and Language Socialization: Ecological Perspectives*, London, New York: Continuum.

Krueger, R A and Casey, M A (2009) *Focus Groups: A Practical Guide for Applied Research*, London: Sage Publications.

Lamprianou, I (2012) Unintended consequences of forced policymaking in high stakes examinations: The case of the Republic of Cyprus, *Assessment in Education: Principles, Policy & Practice* 19 (1), 27–44.

Lamprianou, I and Afantiti Lamprianou, T (2013) *Charting Private Tutoring in Cyprus: A Socio-Demographic Perspective*, in Bray, M, Mazawi, A E and Sultana, R G (Eds) *Private Tutoring Across the Mediterranean: Power Dynamics and Implications for Learning and Equity*, Rotterdam: Sense Publishers, 29–56.

Lantolf, J P (Ed) (2000) *Sociocultural Theory and Second Language Learning*, Oxford: Oxford University Press.

Lantolf, J P and Thorne, S L (2006) *Sociocultural Theory and the Genesis of Second Language Development*, Oxford: Oxford University Press.

Litosseliti, L (2003) *Using Focus Groups in Research*, London: Continuum.

McNamara, T and Roever, C (2006) *Language Testing: The Social Dimension*, Malden: Blackwell Publishing.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13, 241–256.

Norris, J M (2008) *Validity Evaluation in Language Assessment*, Frankfurt: Peter Lang.

Paltridge, B and Phakiti, A (Eds) (2010) *Continuum Companion to Research Methods in Applied Linguistics*, London: Continuum.

Papadima-Sophocleous, S (2013) High-stakes language testing in the Republic of Cyprus, in Tsagari, D, Papadima-Sophocleous, S and Ioannou-Georgiou, S (Eds) *International Experiences in Language Testing and Assessment*, Frankfurt am Main: Peter Lang, 193–217.

Pavlou, P (2010) Preface, in Pavlou, P (Ed) *Research on English as a Foreign Language in Cyprus*, Nicosia: University of Nicosia Press, x–xx.

Rea-Dickins, P (1997) So why do we need relationships with stakeholders in language testing? A view from the UK, *Language Testing* 14 (3), 304–314.

Taylor, L (2000) Stakeholders in language testing, *Research Notes* 2, 2–4.

Tsagari, D (2009) *The Complexity of Test Washback: An Empirical Study*, Frankfurt am Main: Peter Lang GmbH.

Tsagari, D (2012a) FCE-exam preparation discourses: Insights from an ethnographic study, *Research Notes* 47, 36–47.

Tsagari, D (2012b) *The Influence of the Examination for the Certificate of Proficiency in English (ECPE) on test preparation materials*, unpublished Technical Report, SPAAN Fellowship for Studies in Second or Foreign Language Assessment, Cambridge Michigan Language Assessments (CaMLA), Ann Arbor, Michigan, USA.

Wall, D (2005) *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights From Testing and Innovation Theory*, Studies in Language Testing volume 22, Cambridge: UCLES/Cambridge University Press.

Winke, P (2011) Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter, *TESOL Quarterly* 45 (4), 628–660.

## Appendix 1: Exam interview worksheet

Interviewer/s: ________________________________
City: ________________________________
Name of the school: ________________________________

**Teachers**

Teachers' attitudes

1. What do you think of the *Cambridge English: First* exam?
2. What do you like about the exam?
3. What do you not like about the exam?

Teachers' training

4. What training have you received before or while preparing students for the *Cambridge English: First* exam? To what extent, could you say that it has helped you?

**Teaching for Cambridge English: First**

5. How do you normally structure your lesson plans when you prepare for this exam? Does it take long?
6. What are your general goals and objectives when teaching students for *Cambridge English: First*?
7. Does preparation for *Cambridge English: First* influence your teaching?
8. Some teachers use a lot of *Cambridge English: First* materials while preparing students for other exams. Is this the case with you? Is there any particular reason for this?
9. What kind of preparation materials do you use most when you prepare for *Cambridge English: First*? (E.g. comprehension texts, past papers, other?)
10. When do you use them most?
11. Does the answer key help?
12. Which parts of the exam(s) are more difficult for students?
13. When the exam dates approach, does the atmosphere change in the classroom?
14. When do your students take the *Cambridge English: First* exam? (In year 7, 8, 9 for example?)
15. How many hours a week do students do preparation for *Cambridge English: First*?
16. Is one year enough for preparing students for the *Cambridge English: First* exam?
17. If students prepare for *Cambridge English: First*, do they take the exam? If not, would you think it would have been a good idea if they did?

**Other stakeholders' attitudes**

18. What is it that makes schools prepare students for the exam?
19. What do you think parents like about the exam?
20. What do you think parents dislike about the exam?
21. What is the value/recognition of the exam in Cyprus?

**Students**

22. What do students like about the exam?
23. What is it that students dislike about the exam?
24. Why do students fail the exam?
25. Do students or parents ever blame the teacher when students fail to pass the exam?

**Learning**

26. Do you think students benefit from the exam preparation process?
27. Do you think that being a holder of *Cambridge English: First* means being a competent and fluent speaker of the English language?

**Curriculum/Materials**

28. Do you think the materials available in the market are helpful to adequately prepare students for the exam(s)?
29. How do you select a coursebook for exam preparation? What are the criteria you use?
30. What are the materials you use for the *Cambridge English: First* exam?
31. Are they enough? Do you use other resources? If yes, which are they?
32. When do you start working with practice tests?
33. Are the *Cambridge English: First* specifications clear?

# The impact of Cambridge English examinations on institutional change

**ANA MARÍA ROZZI** UNIVERSIDAD CAECE, BUENOS AIRES, ARGENTINA
**VERÓNICA PINTO** UNIVERSIDAD CAECE, BUENOS AIRES, ARGENTINA
**MARINA GONZÁLEZ** UNIVERSIDAD CAECE, BUENOS AIRES, ARGENTINA
**YANINA CRIMI** UNIVERSIDAD CAECE, BUENOS AIRES, ARGENTINA

## Introduction

Cambridge English examinations used as proficiency tests are likely to shed light on learners' and teachers' achievements more objectively than internal examinations, as they measure progress against international standards. Schools which produce their own tests rather than adopt external examinations may remain somewhat unaware of their shortcomings and limitations, as these examinations may be unconsciously tailored to the students' capabilities as perceived by their teachers. When external examinations are adopted, these problems may come to light, and conflict may arise, but of a positive nature, causing the organisation to also change as an institution, that is, to even change its beliefs (North 1990).

Research has often focused on the impact of the adoption of external examinations on teaching and learning, but this impact also affects other aspects of an institution's work and infrastructure. The nature of the relationship between language school management and examinations has received less attention than the pedagogical considerations related to the adoption of external Cambridge English examinations, even though management principles, orientation and vision determine the pedagogy at any school, and the decisions regarding forms and methods of assessment are necessarily taken at management level. Thus, the adoption of external examinations could unveil strong and weak points of teaching and management and serve as a catalyst for institutional change at pedagogic, marketing, administrative and economic levels.

The adoption of external Cambridge English examinations by a language school means introducing an innovation and it should be treated as such (Saville 2009). This normally means that the institution should go through a process of exploration of its previous experience in testing, pre-existing characteristics of the school and needs analysis, followed by the actual implementation and an evaluation of results (White, Martin and Hodge 2002). An innovation, then, is likely to have an impact on the whole of the language institution, causing adaptation and adjustment of its management, curriculum and teaching to the new situation, and the process of evaluation should address not only the learning outcomes and teaching methods, but the macro problem of how the organisation has adjusted, is adjusting or should adjust to the innovation. In this sense, we should no longer talk about *washback* of testing as focused only on learning outcomes and teaching practices, but on the *impact* it may have on management in its broader sense (Nicholls 1983).

At this point, it is pertinent to define some terms. Saville considers impact to be 'a broader notion [of the effects and consequences of tests] operating at both micro and macro levels' (2009:ii). He clarifies that impact is a relatively recent term in language assessment that can be defined as the 'superordinate concept covering the effects and consequences of tests and examinations throughout society' and 'an extension of the notion of washback' (2009:3). The adoption of external exams impacts on the process of teaching, learning, assessment and the development of other aspects such as management, infrastructure, marketing, etc. in language schools. Clear criteria and goals should be the base for measuring success and weak points of the implementation of such tests (Saville 2009).

Although impact and washback are closely related, they are distinct concepts. While the former refers to the results of the adoption of tests in general in institutions, the latter refers to the changes in teaching and learning due to the influence of examinations. White et al (2002) define *organisations* as networks of relationships among the actors who belong to them and state they can be considered as social units that fulfil the needs of their members. The interaction among the individuals will respond to the goals or objectives of the organisation. Not only are language schools organisations, they are also *institutions*, since institutions are driven by 'rules of the game in a hierarchical order' (Aoki 2007:1). Along this line of thought, North (1990) states that institutions consist of the formal norms and informal conventions, and their effectiveness depends on the relationship between institutional goals and participants' personal objectives. He further explains that 'if institutions are the rules of the games, the organisations are the players: groups of individuals engaged in a purposive activity' (1990:3).

How do institutions change? According to Roland (2005), some institutions typically change quickly and others slowly. He explains that there is a difference between institutions depending 'on whether they change slowly and continuously or rapidly and irregularly' (2005:4). He places schools within the first group precisely because they are social instruments for education and as such, they move with social change, which is relatively slow. However, institutions are not completely dependent on external forces and they have the necessary power to change swiftly with a dynamic type of management which will also know how to endogenise external forces (Aoki 2007), that is, sense and appraise social change and adapt to it. The adoption of external examinations may be viewed by some members of the institution as an intrusion into their 'domain' and by others as an opportunity for their organisation to rise to the challenge of achieving international standards of excellence. In both cases, conflicts will arise and they are bound to produce changes in all sectors of an organisation.

According to Roland, *institutional culture*, which encompasses values, beliefs, and social norms, is slow-moving because '[its] evolution is related to the evolution of technology and scientific knowledge' (2005:4). On the contrary, fast-moving institutions can change overnight but this does not mean they are constantly changing. Roland (2005) proposes a classification to understand institutional change based on the ability of institutions to change slowly or rapidly and based on continuous or discontinuous change. Social norms, technology and culture have a tendency to evolve periodically and gradually. But what is the relationship between the two types? The incorporation of international exams can have an impact on institutions which can be a part of any of the mentioned categories of change. The change can be fast when it comes to the incorporation of tests and at the same time slow as regards norms, participants' views, ways of teaching, learning, assessing, staff training, marketing and management.

The implementation of external examinations can produce changes in the school rules and procedures (such as scheduled lessons, instructions on how to prepare syllabuses, or types of materials to be used in each class, among others) and will probably be changed due to the adoption of a tool to evaluate students' performance. Although these rules and procedures generally reflect the 'cultural and ideological belief system of agents' (Aoki 2007:8), they are expected to change with the introduction of an instrument to evaluate students externally. North (1990) believes that change is vital for further progress. Adopting external exams will probably impact institutions in different aspects such as formal and informal rules, objectives and the actors' expectations and needs. Discussion about impact and how it should be managed successfully in the educational system is concerned with the 'interplay between many sub-systems and cultures and the roles of stakeholders as participants [as] a vital factor'

(Fullan 1993, 1999, Thelen and Smith 1994, Van Geert 2007 as cited in Saville 2009:5). Saville (2009:8) relates impact to operational processes, stating that for 'Cambridge ESOL, [it] need[s] to combine theoretical substance with practical applications and to become an integral part of the operational test development and validation framework'.

This brief overview of the literature does not seem to offer a comprehensive analysis of the pedagogic, managerial, cultural and economic impact produced by the adoption of Cambridge English examinations in language schools. It is the main purpose of the present qualitative case study to provide some empirical insights into this previously neglected niche.

## Research objectives

General objective:
To establish the extent and characteristics of institutional change brought about by the adoption of Cambridge English external international examinations at four language schools in Buenos Aires, Argentina. Language schools are defined as establishments where only foreign languages are taught.
Our specific objectives were:

1. To determine why the examinations were adopted.
2. To establish whether the schools changed their syllabuses because of the adoption and how.
3. To find out what training the staff received for preparing learners for the examinations, if any.
4. To compare the marketing actions the school carried out before and after the adoption.
5. To find out whether there is an impact on income that could be attributed to the adoption.
6. To establish whether administrative procedures had to be changed.
7. To analyse possible changes to institutional culture produced by the adoption and those which may occur in the near future.

### Method

This research is a qualitative case study involving four large language schools in Buenos Aires which adopted external examinations offered by Cambridge English Language Assessment.

### Participants

The participants were Heads and some teachers at four language schools. The schools were chosen according to the following criteria:

1. They were preparation centres, not examination centres, i.e. they prepared students to take Cambridge English examinations, they did not manage examinations.
2. They had at least 300 students.
3. They had adopted Cambridge English examinations at least five years before this study.
4. They were all located in Buenos Aires and the Greater Buenos Aires area, i.e. the suburban area surrounding the capital city of Argentina.
5. They enjoyed a reputation of excellence within their communities. (To determine this, the researchers relied on their knowledge of the English teaching situation in the Greater Buenos Aires area.)

For the purposes of respecting anonymity, the schools were numbered from 1 to 4.

- School 1 is located in Greater Buenos Aires. It has a branch in another district as well. It was founded in 1990 and it has 445 students. The school is a member of Schools of English Association (SEA)[1]. Among its services, the school offers courses for children, teenagers, and adults and it organises educational visits to the UK. There are also special preparation courses for Teacher Education Colleges' and Translation Colleges' admission examinations. Cambridge English examinations include *Cambridge English: Young Learners Starters, Movers, Flyers, Cambridge English: Key, Cambridge English: Preliminary, Cambridge English: First, Cambridge English: Advanced* and *Cambridge English: Proficiency*.
- School 2 is located in Buenos Aires. It was founded in 1981 and it has 432 students. The school is a member of SEA. Among its services, the school offers courses for toddlers, children, adolescents, and adults, and short courses and seminars for teachers. The school also offers courses abroad and educational trips to the UK. External examinations include *Cambridge English: First* and *Cambridge English: Advanced*.
- School 3 is located in Greater Buenos Aires. It was founded in 1975. It has 304 students. The school is a member of SEA. Among its services it offers courses for children, adolescents and adults. The school holds courses and seminars for teachers, as well. Courses include outings and guided tours in English and the school organises an annual educational trip to the UK. External examinations include *Cambridge English: First* and *Cambridge English: Advanced*.
- School 4 is located in Greater Buenos Aires. It was founded in 1982 and it has 323 students. The school is a member of SEA. Among its services it offers courses for children, adolescents, and adults. External examinations include *Cambridge English: Preliminary, Cambridge English: First* and *Cambridge English: Advanced*.

At these schools, the research team interviewed the Heads, who were also the owners, as well as four teachers in each school. The teachers interviewed were chosen at random from those who were responsible for preparing students for Cambridge English examinations. A sample of four teachers in each institution was considered sufficient for the purposes of this study. The research team held a preliminary interview with each Head before including the school in this study.

### Data collection

Data collection was based on:

a. Interviews with the Heads and the teachers, held by two researchers. Conversations were recorded and transcribed for analysis.
b. Inspection of the premises, which included taking photographs.

[1] Asociación de Escuelas de Idiomas de la República Argentina

c. Inspection of available documents, such as web pages, syllabuses, examination scores, marketing material, and financial records.

*Instruments*

a. Questionnaires (Appendix 1) to interview the Heads and teachers.
b. Observation checklists (Appendix 2) used to record the data from the inspection of the premises and documents.

## Results

In the interviews with the Heads, their answers to questions 1–6 in the questionnaire (See Appendix 1) revealed that they had decided to adopt Cambridge English examinations to be properly positioned in a market where these examinations are in demand, particularly those intended for children and adolescents. Teachers were not consulted, because Heads considered that consulting their teaching staff about the adoption of the examinations was irrelevant.

These Heads suggested that they would have never chosen a national examinations board, although there are some which enjoy an excellent reputation. Market considerations were again mentioned as one of the reasons, for the Heads sense that the public respects international qualifications more than local ones. It should be pointed out that by 'foreign' parents usually mean 'British', as there is widespread consensus that British English is 'the real English' in some sectors of Argentinian society.

Before adopting the Cambridge English examinations, the schools used their own tests to evaluate students and they continue using them, as not all students sit Cambridge English examinations. At all four schools surveyed, teacher-produced tests are used to promote students from one level to the next, regardless of the results of international examinations. On the other hand, the teachers' assessment is used to advise students on whether to sit international examinations. At three of the schools, students who express their desire to take the examinations are placed in special courses where they get the necessary preparation. In all, approximately 5% of students actually sit the examinations, a fact which also explains the high pass rates these schools boast. One of the schools reported that 25% of their students take Cambridge English examinations, but they are those who enrol for preparation courses or who are advised to sit the exams. All the Heads reported that parents were perfectly happy with this situation and they even suggested that families did not pay any attention to whether their children were chosen to sit the examinations or not.

The Heads' answers to Part 2 in the questionnaire in Appendix 1, information recorded in the checklists in Appendix 2 and the answers to the questionnaire for teachers in Appendix 1 provided information related to other aspects of the impact of the adoption on the schools' work that is discussed below.

When these schools were founded, their syllabuses were tailored to match Cambridge English syllabuses for all levels, from *Cambridge English: Young Learners* to *Cambridge English: Proficiency*. Syllabuses are designed by the Heads, but teachers may suggest modifications and adaptations as the courses develop. As has already been explained, some courses at these schools prepare students specifically to take Cambridge English examinations and others do not. Syllabuses tend to be more rigid for the former and they are usually based on the contents of international textbooks meant to be used for examination preparation. The tendency in Argentina is to use international coursebooks for the teaching of English as a foreign language. Very little local material is produced and it is aimed at a different audience: some state-run schools in the provinces. In the case of international exam preparation, series of textbooks produced abroad are used and Cambridge English offers a variety of support material that is readily available in Buenos Aires. Nevertheless, we were not able to identify exactly what support materials are used at the four language schools under study, as teachers are free to draw them from different sources and no record is kept of the choices they make.

According to teachers and Heads, teaching and learning approaches at these schools can be broadly defined as communication-orientated, both before and after the adoption of Cambridge English examinations. This definition is also present in the schools' websites. However, when Heads were pressed for a more scientific definition of the approach, they were not able to explain how it related to their choice of offering preparation courses for external exams. In one case, vocabulary was cited as the key issue in communicative achievement and in fact, many posters with word webs and similar vocabulary presentations decorated the walls and blackboards of the school in question. Further inspection of classroom friezes showed an emphasis on grammar charts and grammar rules at three of the schools.

Exam preparation was incorporated into the courses leading to a Cambridge English certification, but care was taken not to make examination-passing the focus of teaching, we were told. Heads claimed they expected their students to pass the examinations with very little practice of the examination itself, just as a result of having improved their command of English. This claim seems contradictory with the distinction they make between their ordinary courses and those which lead specifically to Cambridge English examinations.

When asked about changes in assessment, the interviewees explained that most internal assessment is still teacher produced, as it was before the adoption, with few institutional guidelines as regards length or focus. The inspection of these tests revealed some discrepancy with the communicative approach supposedly used for teaching, and it was noticed that some of them had been constructed on the model and format of Cambridge English examinations. In courses devoted to examination preparation, mock Cambridge English examinations are used extensively.

At two schools, the Heads explained that simplified Cambridge English-like examinations are used to promote students to the following level set by the school's internal curriculum. For example, some sections of original mock examinations are eliminated, some are shortened, or some texts are simplified.

Regarding teacher training, either in-house or provided by external advisors or other institutions, the Heads explained that they hire only certified teachers and then offer them some training in fields related to Cambridge English exams. Training sessions are not usually held by local Cambridge English representatives. Heads admitted not having contacted

Cambridge English to enquire whether there was training available for language schools. They resort to their more experienced teachers to coach those who have no experience in exam preparation and use support materials from the Cambridge English website.

Regarding marketing actions, the research team found that large billboards advertising the examinations were displayed inside and outside the schools. Pass rates are not made public in all cases, but bulletin boards inside the school and websites display photos of teachers and successful learners receiving their certificates. All four schools advertise their courses by email and on social networks, and highlight preparation for Cambridge English examinations as one of the school's assets.

It was noteworthy that when asked about their marketing actions, most of the Heads were taken aback by the question and answered that they did not think they engaged in any. They considered that marketing referred exclusively to advertising in the media and had not realised that their communication with parents, the community and other groups, their management style and even the decoration and equipment displayed on their premises could also be marketing actions.

Financially, there has not been a dramatic acceleration in the rate of growth of student numbers and consequently, higher income, since the adoption of Cambridge English exams: it ranged from 10–15% initially, then stayed at 5–10% per year.

As regards administrative work, the only change reported was related to the enrolment of students for the examinations and the receipt and communication of results, for which new procedures had to be created. Further staff were not needed as few students sit the exams.

Changes in equipment and facilities, according to the Heads, were not directly caused by the adoption of Cambridge English examinations. These are schools which enjoy an excellent reputation and pay attention to quality. All rooms are fully equipped in terms of furniture and audio equipment. There are TV sets in some classrooms. No PCs were observed anywhere outside the administrative section in one case, but there is a computer room in each of the other schools.

Institutional culture presented some interesting and distinct characteristics. Given that the Heads are sole proprietors and specialists in English Language Teaching (ELT), and the main decision-makers at their schools, there is a co-existence of a strong personal leadership with a task-orientated culture.

This seems to have remained more or less stable in the past 10 years (2003–13), although the increase in the number of students has led to the appointment of coordinators in some cases. This is actually a result of the adoption of Cambridge English examinations, however, because it helped increase enrolment and thus created the need for more horizontal management, to some extent. Only one of the Heads acknowledged that she had had to appoint a specialist teacher to lead the process of adoption of international examinations.

In the interviews, teachers confirmed they had not had a say in their school's decision to adopt Cambridge English examinations, although at one school one of the teachers had worked in cooperation with the Head to implement the adoption. On the other hand, most teachers surveyed did not expect to be consulted.

Teachers explained that they are used to working at schools where students sit these examinations and therefore do not consider their adoption a real innovation. They all claim to adhere to the communicative approach to teaching, although we have already explained that the tests they use and the friezes in their classrooms show a slightly different approach. Teachers directly involved with examination preparation confirmed they follow the syllabuses in the textbooks, e.g. *Objective First Certificate* (Capel and Sharp 2008) or *Insight into PET* (Naylor and Hagger 2004), published by Cambridge University Press.

Teachers also stated that no significant change had taken place after the adoption and provided the same information as the Heads regarding the design of the syllabuses and the methods used for selecting those students who would sit for the international examinations.

We believed the most interesting question we would pose was, 'Do you find that your evaluation of your students is consistent with the results they obtain in the international examinations? If it is not, what do you do about it?' We wanted to find out, for example, how teachers reacted when some students they considered 'weak' passed the international examinations with good marks and vice versa, and what changes this produced in their methods of teaching and assessment. The question remained unanswered, as candidates are chosen among the 'good' students and specially prepared to pass the Cambridge English examinations. 'Weak' students are advised not to sit and this advice is rarely disregarded, above all because parents do not want to risk paying for examinations their children are likely to fail.

The teachers' answers to questions related to equipment, facilities and administration simply supported the information we had obtained from the Heads and/or observation.

## Conclusions

We undertook to research to what extent the adoption of Cambridge English examinations had produced an institutional change in four language schools, in pedagogic, administrative and managerial areas. Our findings showed that the adoption of Cambridge English examinations did not produce significant institutional change in the schools under study. When these institutions were created, their Heads designed the syllabuses for all the courses on the basis of the recommended syllabuses for Cambridge English examinations. Adopting the examinations was a foreseeable development which neither the Heads nor the teachers considered an important innovation, since they deem it a necessary part of a school's work, for which they were already prepared.

Neither did the adoption produce important changes in assessment methods, as when students fail to reach Cambridge English standards, sometimes syllabuses are altered and special courses are created to cater for lower levels of achievement which continue to be evaluated with teacher-produced examinations, often created on the model of Cambridge English examinations.

The adoption had a stronger impact on the schools' marketing, as it positioned them as quality institutions within their communities. Although examinations are primarily meant to be instruments for assessing students' progress, these schools use them mostly as marketing tools even when some of them claim they attach very little importance to marketing. Candidates are chosen among those who will not fail and are specially prepared for the examinations, so that the school can boast high pass rates. While advertising emphasises the fact that these schools prepare for Cambridge English examinations, very few of their students actually benefit from this preparation. The examinations were adopted because of their prestige and popularity and to be included in the schools' advertising, not to function as instruments to assess achievement or to raise the level of teaching and learning.

Little attention was paid to teacher training, and the situation after the adoption only included coaching of beginner teachers for examination preparation by their more experienced colleagues, but not a comprehensive analysis of needs and the implementation of a training programme in order to improve the overall level of teaching. This may have been considered unnecessary, as the schools opted for preparing only the fastest learners to sit for the Cambridge English examinations.

The schools' marketing orientation, however, proved insufficient and may explain why the adoption produced only a moderate increase in the number of students, as it simply identified the school as one more institution offering Cambridge English examinations preparation, with no distinctive characteristics. The schools did not introduce any innovative methodology or quality-orientated syllabuses to actually enable all their students to reach Cambridge English standards.

According to the Heads' and the teachers' testimonies, language schools fall into two categories in the public's view: good schools which offer international examinations, preferably from Cambridge English, and mediocre schools which use local or internal assessments. When analysed from this perspective, it is clear why neither schools nor parents seem worried by the fact that only 5–25% of students actually sit for the examinations, as examinations are socially viewed only as features of the school's excellence, not as assessment instruments. This outlook might also explain why Heads appear relatively unaware of their marketing actions and of the decisive influence that Cambridge English examinations have had on their syllabuses, test formats and the general orientation of their schools: this is their institutional identity, which they seem to have internalised and taken for granted. Because of this, institutional culture did not change with the adoption and continues to be task-orientated, with a strong leader who makes most of the decisions.

We have said that the adoption was an innovation. The steps in institutionalising an innovation outlined by White et al (2002) were only partially taken: the schools' previous experience in testing was analysed, pre-existing characteristics were considered, needs analysis was carried out, but all these processes were very narrowly focused on the schools' position in the market. As a consequence of this, their impact was felt only in marketing and financial areas and did not constitute a real extension of washback (Saville 2009) involving all the aspects of the institution.

Roland (2005) has already warned us that institutional culture moves slowly and that change can be slow, particularly in pedagogical areas, because it entails changes in people's views, norms and social conventions. Given the necessary agreement outlined by North (1990), White et al (2002) and Aoki (2007) between participants' actions and the goals or objectives of an institution, and the fact that the present conditions seem satisfactory to Heads, teachers, students and parents, no significant changes to the situation outlined in this study seem likely in the near future, particularly in what concerns institutional culture.

## References

Aoki, M (2007) Endogenising institutions and institutional change, *Journal of Institutional Economics* 3, 1–31.

Cambridge English (2013) *Principles of Good Practice: Quality Management and Validation in Language Assessment*, Cambridge: Cambridge English Language Assessment, available online: www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf

Capel, A and Sharp, W (2008) *Objective First Certificate*, Cambridge: Cambridge University Press.

Naylor, H and Hagger, S (2004) *Insight into PET*, Cambridge: Cambridge University Press.

Nicholls, A (1983) *Managing Educational Innovations*, London: Macdonald.

North, D (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.

Roland, G (2005) *Understanding institutional change: Fast-moving and slow-moving institutions*, available online: emlab.berkeley.edu/users/groland/pubs/gr3.pdf

Saville, N (2009) *Developing a model for investigating the impact of language assessment*, available online: http://cambridgeassessment.files.wordpress.com/2011/12/developing-a-model-for-investigating-the-impact-of-language-assessment.pdf

White, R, Martin, M and Hodge, R (2002) *Management in English Language Teaching*, Cambridge: Cambridge University Press.

## Appendix 1: Questionnaires

### Preliminary interviews with Heads

1) Since when has your school been in operation?
2) How many students are there at your school? Has the number been growing in the last few years?
3) What services do you offer?
4) Since when have you been a preparation centre for Cambridge English examinations?

5) What has been the impact of these examinations on your school? (Discuss in very broad terms, to find out whether the Head is actually aware of possible changes).

## Guidelines for interviewers

The following guidelines should help interviewers' work as a memory aid during interviews. However, interviewers should not turn the conversation into a question-and-answer session. During the conversation, one of the interviewers will be taking notes and following the course of the interview, making sure that all the topics are covered, though not necessarily in the order presented in this document. The table and the questions should be used to write the ensuing report.

## Interviews with Heads

### Part 1 – The adoption

1) Did you use any external examinations to evaluate your students' language proficiency before adopting the Cambridge English exams?
   1.1. If so, which?
   1.2. If so, why did you change?
   1.3. Who was involved in making the decision to change/select Cambridge English exams?
2) If you did not use any external examinations, how did you test your students?
3) What are the expected impacts/goals for using Cambridge English exams?
   a. on students?
   b. on teachers?
   c. on other stakeholders?
4) Why did you choose a foreign examinations board instead of a national one?
5) Who did you consult before introducing the tests?
6) What was the teachers' reaction to the change?

### Part 2 – Changes

For each item, explain the situation before and after the adoption. The questions are to be used as guidelines.

| | |
|---|---|
| **1. Syllabus design and designers** | |
| Who was in charge of designing syllabuses before the adoption? | Were any changes necessary after the adoption in terms of<br>a. Objectives<br>b. Coursebooks used<br>c. Emphasis on speaking skills<br>d. Emphasis on writing skills<br>e. Emphasis on listening skills<br>f. Emphasis on reading skills<br>g. Reflecting on one's performance and test taking strategies |
| As regards teaching/learning approaches | |
| | Have you perceived any increase/decrease (Inc/dec) in teachers' work in class in the areas of:<br>1. Amount of L2 spoken Inc/dec<br>2. Peer and group interaction Inc/dec<br>3. Amount of speaking activities Inc/dec<br>4. Amount of writing practice Inc/ dec<br>5. Amount of specific test training inc/ dec |
| As regards assessment | |
| What kind of assessment did the organisation apply before the adoption?<br>1. Teacher-led<br>2. Level-led<br>3. Institutionally defined<br>a. Competence-oriented<br>b. Content-oriented<br>c. Grammar performance-oriented<br>d. Skills-oriented | And after the adoption?<br>Did the role of internal assessment change?<br>Did teacher interaction/sharing change in exam design?<br>Did Cambridge English tests affect the design of internal assessment? How? |

| | |
|---|---|
| **2. Teacher training offered by the school, if any. (Either in service or through external courses or trainers)** | |
| **Before the adoption** | **After the adoption** |
| Did the school have a staff training programme in place before the adoption?<br>Did it have a set number of sessions per year? | Did it change?<br>How did it change?<br>Did teachers receive special training by Cambridge representatives?<br>If so, when was that? |
| **3. Marketing actions** | |
| **Before the adoption** | **After the adoption** |
| What lines of communication were in place before the adoption (with parents, students, staff, the public)? | How did the school communicate the innovation to the parents and students?<br>Were there any changes in the lines of communication? If so, which? |
| **4. Income and rate of growth** | |
| **Before the adoption** | **After the adoption** |
| Did the organisation show a pattern of income growth? Was it steady? | Were any changes in the growth pattern perceivable? If so, were they positive or negative? In what ways? |
| **5. Standard administrative procedures** | |
| **Before the adoption** | **After the adoption** |
| Can you briefly describe the registration process in your institution before the adoption in terms of frequency, amount of data gathered from students and staff devoted to the task? | Were any changes perceivable in the frequency, data and staff devoted to enrolment? |
| **6. Institutional culture** | |
| **Before the adoption** | **After the adoption** |
| Did teachers have a room where they can share time, experiences and materials?<br>Did teachers come to you with suggestions on teaching or on materials or students' needs?<br>Were there any specialists in a skill or level?<br>Were new teachers introduced with the help of senior members?<br>Were teachers assigned group tasks with some peers? Can you expand?<br>What kind of relationship was there between teaching and administrative staff? | Have any changes in the behaviours mentioned before been perceived?<br>Could you explain? |
| **7. Equipment and facilities** | |
| **Before the adoption** | **After the adoption** |
| Briefly describe the rooms you had available:<br>Equipment for teachers and teaching<br>Equipment for administration<br>Library facilities<br>Online resources | What changes, if any, took place after the adoption in the areas mentioned before? |

## Interview with the teachers

**The adoption**

1) Were you involved in the decision to adopt the Cambridge English tests?
2) What was your reaction to the change? Were you happy with the decision to introduce the exams?
3) Do you think the exams are a useful measure of learners' language ability?
4) What is language learning for you?
5) Why, do you think, is your school using these tests?
6) What do you think are the goals/impacts of using these tests on:
   a. Students?
   b. Teachers?
   c. Parents?
   d. The school?
7) Have you observed any changes since the introduction of the exams?

8) Did the syllabuses change after the adoption?
9) Did your teaching strategies change after adoption?
10) Did your assessment strategies change? Did you take part in the design of the new syllabuses?
11) Do these examinations contribute to your motivation?
12) Do these examinations contribute to students' motivation?
13) Do you find that your evaluation of your students is consistent with the results they obtain in the international examinations? If it is not, what do you do about it? For example: learners who have been scoring very high marks during the school year then fail the external examination and vice-versa. How do you use this feedback?
14) What kind of training have you received for preparing students for these exams?
15) Are parents interested in their children taking these exams? Why? Why not?
16) Is this institution growing steadily?
17) How do you communicate with:
    a. Parents?
    b. Other teachers?
    c. The head/manager?
18) Do you help in administrative work?
19) Do you find the facilities comfortable and adequate?
20) Is there enough equipment for you to use different resources, such as web-based activities, videos, books, etc.?

## Appendix 2: Observation checklist and results (sample answers)

| **Observation checklist** | | | | |
|---|---|---|---|---|
| **Dimensions** | Aspects | Document available | Comments | |
| | | Yes | No | |
| **INSTITUTION** | Mission | | X | |
| | Structural organisation | X | | On the web pages. |
| | External and internal communications | X | | |
| | Use of technology | X | | Visual inspection of the classrooms. |
| | Sufficient infrastructure | X | | Visual inspection. |
| **PARTICIPANTS** | | | | |
| **HEAD/OWNER** | University graduate with a degree in TEFL | X | | |
| | Deeds, contracts or other proof of ownership | X | | |
| **MARKETING** | Bulletin board | X | | |
| | Web page | X | | |
| | Leaflets | X | | |
| | Facebook | X | | |
| | Library | X | | |
| | Online news | X | | |
| | Emails | X | | |
| **DOCUMENTS** | Access to docs | X | | Not to all documents. Some were not available on the premises, e.g. balance sheets. |
| | Results of exams | X | | On the web pages. |
| **CLASSROOMS** | Blackboards | | X | |
| | Audio equipment | | X | |
| | Friezes | | X | |
| | Suitable furniture | | X | |
| | Video equipment | | X | |
| | Computers | X | | Computers were in the computer lab, not in the classrooms. |

# Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels

**OKIM KANG** NORTHERN ARIZONA UNIVERSITY, USA
**LINXIAO WANG** NORTHERN ARIZONA UNIVERSITY, USA

## Introduction

In the area of second language (L2) assessment of speaking ability monologic tasks are widely used. Meanwhile, interactive tasks have been receiving increasing attention in language assessment. In general, researchers agree that different task types can affect test takers' performance patterns in the domain of speech styles, lexis and grammar (Kim 2009, Skehan 2001). While much research thus far has examined the relationship between linguistic features in oral performances and proficiency levels (Brown, Iwashita and McNamara 2005, Kang 2013), studies on how different task types affect linguistic features in speaking assessment are rare. Some studies (e.g., O'Sullivan, Weir and Saville 2002) focused on language functions across task types, but linguistic features elicited via individual and paired tasks have not been compared to a great extent. In addition, it is not clear what linguistic features are involved in candidates' performances in their interactive conversation and how these features contribute to the distinction of their proficiency levels.

Therefore, the current article examines the impact of task types on candidates' spoken responses and interactive features that can distinguish proficiency levels. The first goal of the current project is to compare the linguistic features of candidates' output in two tasks: (1) a long turn task (the individual monologue task) and (2) a paired task which involves a two-way conversation between two candidates (the interactive task). The tasks are part of the Speaking tests in the following examinations: *Cambridge English: Preliminary, Cambridge English: First, Cambridge English: Advanced*, and *Cambridge English: Proficiency*. (*Cambridge English: Key* has been excluded from this project because it does not contain a long turn task.) The second goal of the current project is to further analyze the interactive tasks to identify the similarities and differences across proficiency levels in terms of discourse-based communicative features, i.e. co-operation, coherence, turn-taking, and strategies.

## Research questions

The project is guided by the following research questions:

1. Are the linguistic features of candidates' speech elicited by individual tasks different from those in interactive tasks at each of the Cambridge English exams' levels (B1–C2) for the following scoring criteria: (a) discourse management (among which this study focuses on the fluency aspect only), (b) grammatical resources, (c) lexical resources, and (d) pronunciation?
2. What are the salient interaction features that distinguish the Cambridge English exams' levels (B1–C2) for the criterion of discourse-based communication: (a) co-operation, (b) coherence, (c) turn-taking, and (d) strategies?

## Methodology

The study applied a quantitative, corpus-based approach to quantitatively analyse data, which is speech samples received from Cambridge English Language Assessment which included spoken responses of both the long turn task and paired task. This paper reports on two studies: Kang (2013), and Kang and Wang (2013). The former investigated linguistic features of the individual task while the latter explored linguistic and communicative features of the interactive task and compared language from the two types of tasks.

### Materials

In total 58 video files provided by Cambridge English Language Assessment were analysed (originally the researchers received 61 videos but three were not analysable due to extra noises and poor sound quality), which included both the individual and interactive tasks. There were 28 tests from *Cambridge English: Preliminary*; 32 from *Cambridge English: First*; 34 from *Cambridge English: Advanced*; and 22 from *Cambridge English: Proficiency*. This translated into 116 individual tasks and 58 interactive ones. For the purpose of this study, 1 minute was extracted from each individual task and 2 minutes from each interactive task. Table 1 provides a breakdown of the two tasks across four Common European Framework of Reference (CEFR) levels.

**Table 1: A breakdown of the two tasks across four CEFR levels**

| Level | Number of individual tasks | Number of interactive tasks |
|---|---|---|
| **B1, Preliminary** | 28 | 14 |
| **B2, First** | 32 | 16 |
| **C1, Advanced** | 34 | 17 |
| **C2, Proficiency** | 22 | 11 |

### Data coding

Individual tasks (1 minute per task x 116 individual interlocutors = 116 minutes) were transcribed and coded for errors or other linguistic features for each of the four scoring criteria in the areas of discourse management, grammatical resources, lexical resources, and pronunciation. Two minutes of each interactive task were extracted so that each speaker in interactions would have approximately 1 minute of speech, which could be comparable with those of individual tasks. The interactive tasks (2 minutes per task x 58 interactions = 116 minutes) were then transcribed and coded for the same four scoring criteria. Interactive tasks were further manually coded for interaction features, which included the following four parts: (a) co-operation, (b) discourse markers, (c) turn-taking, and (d) conversation maintenance strategies. Coding was conducted both manually and automatically, using PRAAT and CSL computer programs for pronunciation and fluency, and VocabProfiler (www.lextutor.ca/vp/eng/) for lexical resource. Inter-coder reliabilities for manual coding were tested and proved to be acceptable (.75 and higher).

### Linguistic analysis

The transcribed interactive tasks were firstly analyzed for individual linguistic features across the four scoring criteria and then for interaction features. Certain features of discourse management and pronunciation were excluded from the current analysis. Coherence features as part of discourse management were not included because the coherence proved to be not comparable between the two tasks. In addition, some of the tone choice variables (e.g. pitch height) were not included, due to different pitch and intonation patterns among interlocutors.

*Fluency (as part of discourse management)*. As for fluency measures, the study examined speech rate measured in other studies (e.g. Kang 2008, Kormos and Denes 2004, Riggenbach 1991), and pause structures of the responses (e.g. Brown and Yule 1983, Kang 2008). We included the following fluency features: (a) syllables per second; (b) mean length of run (*run* is defined as utterances between pauses of 0.1 seconds); (c) phonation time ratio (i.e. the percentage of time spent speaking/total time taken to produce the speech sample); (d) total number of silent pauses; (e) mean length of silent pauses; (f) total number of filled pauses; and (g) mean length of filled pauses.

*Lexical resource*. The use of the lexicon has been measured through vocabulary richness and range (e.g. Brown et al 2005). Vocabulary richness was calculated as a proportion of low and high frequency vocabulary used in each spoken response. Vocabulary range was measured by type-token ratio (TTR). For lexical resources, there were (a) total number of types (different words); (b) total number of tokens (words in text); (c) total number of K1 tokens (the most frequent 1,000 words of English); (d) total number of K2 tokens (the second most frequent thousand words of English i.e. 1,001–2,000); (e) total number of academic word list (AWL) tokens; (f) TTR (a ratio of the number of different words to the number of total words); (g) lexical density (the number of content words divided by total number of words); and (h) total number of word families. The TTR is used as a measure of lexical diversity.

*Grammatical resource*. This criterion includes grammatical complexity and accuracy. Grammatical complexity was measured through verb–phrase ratio and occurrences of grammatical features. Grammatical accuracy measures included global accuracy (Brown et al 2005) and specific types of errors (e.g. tense marking, plural, preposition) (Iwashita, Brown, McNamara and O'Hagan 2008). Global accuracy was measured through error-free T-units. A T-unit is defined as an independent clause and all its dependent clauses (Hunt 1970). Error-free T-units are T-units free from any grammatical errors. In the analysis of grammatical complexity and accuracy, we included (a) total number of T-units; (b) total number of error-free T-units; (c) T-unit complexity; (d) total number of clauses; (e) total number of dependent clauses; and (f) specific types of errors including the use of: tense marking; singular/plural, prepositions, articles, adverbs, pronouns, adjectives, verbs, determiners, subjects, objects, negators, copulas, modals, relative clauses, non-finite clauses, and passive.

*Pronunciation*. The focuses of analysis were stress and pitch. The variables selected are accented measures in pronunciation called 'acoustic fluency', and are the best predictors of rated oral performance (Kang 2008, 2010, Kang, Rubin and Pickering 2010). Therefore, the project included the following variables for Pronunciation measures: (a) tone choice (falling); (b) tone choice (level); (c) tone choice (rising); (d) pitch range; (e) pace (the average number of prominent syllables); and (f) space (the proportion of prominent words to the total number of words). The measures from (a) to (d) were elements of pitch and intonation, and the last two measures, (e) and (f), were exponents of stress. Note that segmental features were excluded from analysis due to low correlations found between segmental errors and listeners' judgments (e.g. Anderson-Hsieh, Johnson and Koehler 1992).

*Interaction features*. Cambridge English tests measure test takers' communicative competence through comprehensive assessment categories. One important aspect is *interactive communication*. Interaction skills are assessed by criteria such as initiating and responding, contributing to conversation development actively, and also using interactive strategies to maintain and repair communication. The current project utilized Celce-Murcia, Dörnyei and Thurrell's (1995) framework for discourse and strategic competence to operationalize this construct. The framework is based on communicative language learning, which is in line with the understanding of *interactive communication* in Cambridge English exams. Variables to measure included interlocutors' co-operation, coherence, turn-taking and strategy use.

Measures of co-operation included variables of back-channelling (the occurrence of utterances briefly responding to a partner in either nonverbal units or phrasal ones), topic initiation (the occurrence of utterances starting a new idea), and overlap initiation (the occurrence of two interlocutors starting their utterances at the same time) in conversations employing a conversation analysis approach (Schegloff 1982). Discourse markers were used as the measure of coherence. They are words or phrases that are relatively syntax-independent and do not change the meaning of the sentence. Regarding the turn-taking measures, the study followed Crookes' (1990) definition of a 'turn' which is 'one or more streams of speech bounded by speech of another, usually an interlocutor' (1990:185) to operationalize turn-takings. With regard to turn-lengths, the study adopted

Gnisci and Bakeman's methods (2007) to operationalize turn lengths as short (1–10 words), middle (11–30 words), and long (more than 30 words). Finally, with regard to the measure of interaction strategies, Nakatani's (2010) response maintenance (utterances mention a part of previous utterances of other interlocutors) and repair strategies (utterances correct one interlocutor's own speech) were utilized. Overall, the study included the following interaction features: back-channelling, prompting, topic initiation, overlapping initiation, discourse markers, total talking time, turn-taking time, number of total turns, total number of short turns (one to 10 words), total number of middle turns (11 to 30 words), total number of long turns (more than 30 words), response maintenance (utterances referring to the other interlocutor's previous utterances), and repair (utterances of self-correcting).

*The analysis of speech by task and proficiency levels*. With regard to the first research question about differences in linguistic features between the individual tasks and interactive tasks at each of the four levels (B1–C2) of Cambridge English Language Assessment examinations, descriptive statistics of linguistic variables were reported by the two task types and four proficiency levels. Two types of paired *t*-tests were conducted to determine if the differences in discourse elicited by the two tasks are statistically significant. Firstly, an overall paired *t*-test analysis was performed to investigate any differences between the two tasks in terms of each of the four scoring criteria regardless of proficiency levels. Then, in a post hoc analysis, linguistic features were submitted to paired *t*-tests again to examine differences within each proficiency level.

To answer the second research question about the salient interaction features which may distinguish B1–C2 CEFR levels, descriptive statistics of each interaction feature were reported by two task types and four proficiency levels. In addition, a series of ANOVA tests were conducted followed by Tukey tests as post hoc analyses to identify the interaction features which exhibited significant mean differences across the four proficiency levels.

## Results

### Discourse management: Fluency

As shown in the Total column in Table 2, four variables demonstrated statistically significant differences between the two tasks in the fluency measures across the levels of proficiency: syllable per second, *t* (115) = 2.05, *p* = .009; mean length of silent pauses, *t* (115) = −20.96, *p* = .000; number of filled pauses, *t* (115) = 3.90, *p* = .000; and mean length of filled pauses, *t* (115) = −5.35, *p* = .000. The results suggested that the candidates in the interactive task spoke faster, paused shorter (both silent and filled pauses), but hesitated longer than in the individual task. However, the test takers in general had more filled pauses (hesitation markers such as 'um' or 'eh') in the interactive task than in the individual long run task.

Paired *t*-tests were performed to see how fluency features differed by task types at each proficiency level. In *Cambridge English: Preliminary*, there were significant differences (p< .01) in the frequency of three variables between the two tasks: syllable per second, *t* (27) = 7.55, *p* = .000; mean length of silent pauses, *t* (27) = −11.36, *p* = .000; and mean length of filled pauses, *t* (27) = −4.30, *p* = .000. That is, the task type affected most of the fluency features. Specifically, examinees at *Cambridge English: Preliminary* produced more syllables per second with more silent and filled pauses in interactive tasks. Meanwhile, examinees at this level had shorter pauses in interactive tasks. Three variables were found to be significantly different between the two tasks in *Cambridge English: First*: number of silent pauses, *t* (31) = −4.02, *p* = .000; mean length of silent pauses, *t* (31) = −11.67, *p* = .000; number of filled pauses, *t* (31) = 3.61, *p* = .001. The task type affected the use of pause at this level; examinees at this level produced fewer and shorter silent pauses but more filled pauses in the interactive task. In *Cambridge English: Advanced*, four features emerged to be statistically significantly different between the two tasks: syllable per second, *t* (33) = 6.37, *p* = .000; mean length of run, *t* (33) = −3.00, *p* = .005; mean length of silent pauses, *t* (33) = −11.97, *p* = .000; and mean length of filled pauses, *t* (33) = −2.92, *p* = .006. At this

**Table 2: Fluency features identified by proficiency levels and task types**

| Features | Task | B1, Preliminary (N = 28) | B2, First (N = 32) | C1, Advanced (N = 34) | C2, Proficiency (N = 22) | Total (N = 116) |
|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Syllable per second*** | Int.** | 2.56 (.66) | 3.01 (.57) | 3.35 (.62) | 3.29 (.60) | 3.06 (.68) |
| | Ind. | 1.84 (.38) | 3.06 (2.88) | 2.64 (.55) | 3.11 (.39) | 2.65 (1.62) |
| **Mean length of run** | Int. | 3.12 (.81) | 3.57 (1.16) | 4.04 (1.05) | 3.51 (.92) | 3.59 (1.05) |
| | Ind. | 3.04 (.62) | 4.18 (3.06) | 4.73 (1.31) | 7.90 (1.06) | 4.77 (6.47) |
| **Phonation time ratio** | Int. | .63 (.13) | .73 (.09) | .71 (.10) | .75 (.07) | .70 (.11) |
| | Ind. | .67 (.08) | .70 (.08) | .72 (.09) | .76 (.06) | .71 (.09) |
| **Number of silent pauses** | Int. | 35.03 (6.49) | 31.65 (8.15) | 29.93 (7.31) | 33.18 (6.22) | 32.25 (7.34) |
| | Ind. | 31.41 (7.10) | 39.01 (8.79) | 32.11 (7.84) | 32.63 (7.21) | 33.95 (8.36) |
| **Mean length of silent pauses*** | Int. | .29 (.09) | .22 (.07) | .24 (.07) | .21 (.07) | .24 (.08) |
| | Ind. | .69 (.21) | .49 (.13) | .57 (.17) | .46 (.13) | .56 (.18) |
| **Number of filled pauses*** | Int. | 8.21 (5.63) | 6.63 (4.26) | 5.62 (4.82) | 5.69 (3.56) | 6.54 (4.73) |
| | Ind. | 5.90 (3.02) | 3.99 (2.62) | 4.82 (3.73) | 4.93 (2.96) | 4.87 (3.17) |
| **Mean length of filled pauses*** | Int. | .07 (.05) | .05 (.05) | .05 (.04) | .04 (.03) | .05 (.01) |
| | Ind. | .14 (.11) | .06 (.05) | .08 (.06) | .08 (.04) | .09 (.01) |

*Note.* * *p*<.01 for overall analysis.

** *Int.* represents interactive tasks. *Ind.* represents individual tasks.

level, candidates produced more syllables per second, but had shorter runs with shorter silent and filled pauses in the interactive task. Finally, in *Cambridge English: Proficiency*, only two features were found to be statistically significant between the two tasks: mean length of silent pauses, *t* (21) = −9.13, *p* = .000; and mean length of filled pauses, *t* (21) = −3.32, *p* = .003. This result suggested that advanced examinees paused differently in length between the two tasks, specifically, making shorter silent and filled pauses in the interactive task.

### Grammatical analysis results

As shown in the Total column in Table 3, four grammatical complexity variables showed statistically significant differences in task types across the four proficiency levels (*p* < .01): total number of T-units in the interactive task, *t* (115) = 10.25, *p* = .000; total number of error-free T-units, *t* (115) = 8.25, *p* = .000; T-units complexity, *t* (115) = −4.24, *p* = .000; and total number of clauses, *t* (115) = 6.02, *p* = .000. These results showed that more T-units, error-free T-units, and clauses were produced in the interactive task than in the individual task. However, candidates across proficiency levels in general (except for B1 level) tended to produce a significantly higher-level grammatical complexity in the individual task than in the interactive task. Overall, when test takers interacted with others, they were inclined to speak more accurately in their grammatical use but used somewhat less complex sentences, compared to a situation where they speak alone.

Paired-sample *t*-tests were also conducted at each proficiency level. In *Cambridge English: Preliminary*, there were significant differences in the frequency of four grammatical complexity variables between the two tasks: total number of T-units, *t* (27) = 3.31, *p* = .003; T-unit complexity, *t* (27) = 3.10, *p* = .005; total number of clauses, *t* (27) = 5.89, *p* = .000; and total number of dependent clauses, *t* (27) = 5.55, *p* = .000. Specifically, in the interactive task, examinees in *Cambridge English: Preliminary* produced more T-units, error-free units, clauses, and dependent clauses. At this level, the grammatical complexity in conversation was significantly higher than that in monologues, which is not the case in the other levels of proficiency. There were significant differences in the frequency of three variables between the two tasks in *Cambridge English: First*: total number of T-units, *t* (31) = 6.32, *p* = .000; total number of error-free T-units, *t* (31) =5.09, *p* = .000; and total number of clauses, *t* (31) = 4.67, *p* = .000. According to these results, the task type affected the total number of T-units, error-free T-units, and total number of clauses significantly at this proficiency level. Examinees at this level produced more T-units, error-free units, and clauses in the interactive task. In *Cambridge English: Advanced*, significant differences were found in four grammatical complexity variables between the two tasks: total number of T-units, *t* (33) = 5.32, *p* = .000; total number of error-free T-units, *t* (33) =5.85, *p* = .000; T-unit complexity, *t* (33) = −3.56, *p* = .001; and total number of clauses, *t* (33) = 2.40, *p* = .022. The results suggest that in the interactive task, examinees at this level produced more T-units, error-free units, and clauses. However, examinees demonstrated a significantly higher grammatical complexity in the individual task rather than in the interactive task. Lastly, results for the *Cambridge English: Proficiency* test showed significant differences between the two tasks in the following variables: total number of T-units, *t* (21) = 6.05, *p* = .000; total number of error-free T-units, *t* (21) = 4.21, *p* = .000; T-unit complexity, *t* (21) = −6.46, *p* = .000; and total number of dependent clauses, *t* (21) = −4.30, *p* = .000. In the interactive task, examinees at this level produced more T-units and error-free units. It is, however, important to note that examinees demonstrated a significantly higher level of grammatical complexity and more dependent clauses in the individual task as compared to those in the interactive task.

The Total column in Table 4 suggested that the overall comparisons between the two tasks demonstrated five statistically significant differences in grammatical accuracy variables: total number of adjectives errors, *t* (115) = 2.90, *p* = .004; total number of verb errors, *t* (115) = 2.97, *p* = .004; total number of relative clauses errors, *t* (115) = −3.20, *p* = .002; total number of determiner errors, *t* (115) = −3.44, *p* = .001; and total number of object errors, *t* (115) = 2.92, *p* = .004. Generally speaking, examinees made more errors in using adjectives, verbs, subject and object in the interactive task than in the individual task. However, significantly fewer errors were found with regard to relative clauses and determiners in conversation.

**Table 3: Grammatical complexity features identified by proficiency levels**

| Features | Task | B1, Preliminary (N = 28) | B2, First (N = 32) | C1, Advanced (N = 34) | C2, Proficiency (N = 22) | Total (N = 116) |
|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Total number of T-units*** | Int.** | 10.12 (3.91) | 13.06 (4.50) | 11.77 (3.04) | 14.97 (3.68) | 12.33 (4.12) |
| | Ind. | 7.57 (1.97) | 8.47 (2.69) | 8.29 (2.63) | 9.00 (2.60) | 8.30 (2.51) |
| **Total number of error-free T-units*** | Int. | 4.34 (3.24) | 6.89 (3.89) | 7.56 (2.83) | 9.71 (3.76) | 7.01 (3.83) |
| | Ind. | 3.04 (2.03) | 2.97 (2.26) | 3.74 (2.69) | 4.73 (3.30) | 3.54 (2.62) |
| **T-unit complexity*** | Int. | 1.79 (.43) | 1.77 (.57) | 1.79 (.38) | 1.72 (.35) | 1.77 (.44) |
| | Ind. | 1.51 (.33) | 2.02 (.70) | 2.20 (.53) | 2.89 (.65) | 2.11 (.73) |
| **Total number of clauses*** | Int. | 17.54 (6.13) | 21.34 (5.38) | 20.58 (4.83) | 25.03 (5.30) | 20.90 (5.87) |
| | Ind. | 11.46 (3.96) | 16.19 (4.78) | 17.74 (5.46) | 25.00 (5.58) | 17.17 (6.64) |
| **Total number of dependent clauses** | Int. | 7.08 (3.37) | 7.76 (4.10) | 8.78 (3.68) | 10.06 (4.17) | 8.33 (3.92) |
| | Ind. | 3.57 (2.64) | 7.38 (4.53) | 8.97 (4.30) | 15.36 (3.75) | 8.44 (5.51) |

*Note.* * *p*<.01 for overall analysis.

** *Int*. represents interactive tasks. *Ind*. represents individual tasks.

**Table 4: Grammatical accuracy features identified by proficiency levels**

| Features | Task | B1, Preliminary<br>(N = 28) | B2, First<br>(N = 32) | C1, Advanced<br>(N = 34) | C2, Proficiency<br>(N = 22) | Total<br>(N = 116) |
|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| Tense errors | Int.** | .61 (1.29) | .17 (.37) | .17 (.43) | .59 (1.00) | .36 (.84) |
| | Ind. | .75 (.80) | .38 (.83) | .65 (.92) | .50 (.67) | .57 (.83) |
| Singular/plural errors | Int. | .93 (.89) | .39 (.60) | .48 (.78) | .63 (.89) | .59 (.80) |
| | Ind. | .71 (.90) | .81 (1.00) | .35 (.54) | .36 (.73) | .57 (.83) |
| Preposition errors | Int. | 1.77 (1.76) | 1.84 (1.40) | 1.89 (4.02) | 1.36 (1.43) | 1.75 (2.51) |
| | Ind. | .68 (.86) | 1.72 (1.22) | 1.56 (1.05) | 1.18 (1.05) | 1.32 (1.12) |
| Article errors | Int. | 1.30 (1.76) | 1.41 (1.33) | .67 (.86) | 1.10 (1.10) | 1.11 (1.31) |
| | Ind. | 1.64 (1.52) | 1.97 (2.01) | 1.00 (1.07) | .86 (1.12) | 1.40 (1.55) |
| Adverb errors | Int. | .35 (.74) | .29 (.45) | .23 (.57) | .26 (.62) | .28 (.59) |
| | Ind. | .32 (.55) | .25 (.51) | .29 (.52) | .32 (.57) | .29 (.53) |
| Pronoun errors | Int. | .31 (.56) | .26 (.61) | .20 (.46) | .28 (.50) | .26 (.53) |
| | Ind. | .29 (.53) | .38 (.71) | .41 (.78) | .55 (.67) | .40 (.68) |
| Adjective errors* | Int. | .46 (.81) | .39 (.73) | .32 (.68) | .24 (.59) | .36 (.70) |
| | Ind. | .11 (.42) | .25 (.57) | .09 (.29) | .09 (.29) | .14 (.41) |
| Verb errors* | Int. | .55 (.70) | .39 (.63) | .41 (.73) | .49 (.81) | .46 (.71) |
| | Ind. | .29 (.53) | .09 (.30) | .21 (.48) | .32 (.65) | .22 (.49) |
| Determiner errors* | Int. | .20 (.45) | .18 (.38) | .05 (.19) | .00 (.00) | .11 (.32) |
| | Ind. | .43 (.57) | .28 (.52) | .26 (.57) | .27 (.63) | .31 (.56) |
| Subject errors | Int. | .48 (.77) | .33 (.55) | .38 (.74) | .39 (.60) | .39 (.67) |
| | Ind. | .18 (.39) | .31 (.82) | .18 (.46) | .05 (.21) | .19 (.54) |
| Object errors* | Int. | .31 (.60) | .22 (.50) | .27 (.50) | .23 (.56) | .25 (.53) |
| | Ind. | .04 (.19) | .13 (.34) | .06 (.24) | .14 (.35) | .09 (.28) |
| Negator errors | Int. | .15 (.51) | .13 (.36) | .00 (.00) | .00 (.00) | .07 (.32) |
| | Ind. | .00 (.00) | .06 (.25) | .03 (.17) | .00 (.00) | .03 (.16) |
| Copula errors | Int. | .64 (.96) | .50 (.73) | .20 (.48) | .40 (.69) | .43 (.74) |
| | Ind. | .32 (.67) | .38 (.71) | .38 (.60) | .09 (.29) | .31 (.61) |
| Modal errors | Int. | .03 (.15) | .07 (.24) | .06 (.23) | .04 (.17) | .05 (.20) |
| | Ind. | .00 (.00) | .03 (.18) | .03 (.17) | .23 (.43) | .06 (.24) |
| Relative clause errors* | Int. | .23 (.45) | .48 (.81) | .24 (.48) | .54 (.89) | .36 (.67) |
| | Ind. | .46 (.74) | .94 (.98) | .62 (.78) | .59 (.80) | .66 (.84) |
| Non-finite clause errors | Int. | .48 (.88) | .62 (1.03) | .10 (.35) | .10 (.32) | .34 (.76) |
| | Ind. | .25 (.59) | .44 (.72) | .12 (.33) | .23 (.43) | .26 (.55) |
| Passive errors | Int. | .03 (.16) | .06 (.26) | .03 (.16) | .00 (.00) | .03 (.18) |
| | Ind. | .11 (.57) | .03 (.18) | .15 (.44) | .14 (.35) | .10 (.40) |

*Note.* * $p<.01$ for overall analysis.

** *Int.* represents interactive tasks. *Ind.* represents individual tasks.

Paired-sample *t*-tests at the B1 level showed that there were three significant differences found between the two tasks: number of preposition errors, *t* (27) = 2.95, *p* = .006; number of adjective errors, *t* (27) = 2.10, *p* = .045; and number of object errors, *t* (27) = 2.23, *p* = .034. These results mean that examinees at *Cambridge English: Preliminary* had significantly more errors in using prepositions, adjectives, and objects in the interactive task than in the individual task. At the B2 level, two variables showed significant differences between the two tasks: number of verb errors, *t* (31) = 2.28, *p* = .030; and number of relative clause errors, *t* (31) = −2.12, *p* = .042. Examinees at *Cambridge English: First* had significantly more verb errors but fewer relative clauses errors in the interactive task, compared to their uses in the individual task. At the C1 level, four variables yielded significant differences between the two tasks: number of tense errors, *t* (33) = −2.67, *p* = .012; number of determiner errors, *t* (33) = −2.18, *p* = .037; number of object errors, *t* (33) = 2.06, *p* = .047; and number of relative clause errors, *t* (33) = −2.66, *p* = .012. That said, candidates at *Cambridge English: Advanced* had significantly more object use errors but fewer errors of tense, determiner and relative clauses in the interactive task than in the individual task. At the C2 level, only two variables indicated significant differences: number of subject errors, *t* (21) = 2.44, *p* = .024; and number of copula errors, *t* (21) = 2.30, *p* = .032. Examinees at *Cambridge English: Proficiency* had significantly more errors of using subject and copula in the interactive task than in the individual task.

### Lexical analysis results

Across proficiency levels, candidates showed statistically significant differences among the following six variables in Table 5 (see the Total column, $p < .01$): total number of types, *t* (115) = 7.61, *p* = .000; total number of tokens, *t* (115) = 4.13,

**Table 5: Lexical features identified by proficiency levels**

| Features | Task | B1, Preliminary (N = 28) | B2, First (N = 32) | C1, Advanced (N = 34) | C2, Proficiency (N = 22) | Total (N = 116) |
|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Total number of types*** | Int.** | 59.95 (18.20) | 70.94 (11.52) | 75.48 (10.66) | 84.71 (17.56) | 72.23 (16.51) |
| | Ind. | 44.36 (9.33) | 60.06 (10.93) | 65.15 (12.55) | 81.82 (9.84) | 61.89 (16.44) |
| **Total number of tokens*** | Int. | 113.64 (33.50) | 139.16 (23.21) | 141.55 (22.36) | 164.23 (29.07) | 138.46 (31.41) |
| | Ind. | 84.50 (22.65) | 125.84 (25.67) | 130.85 (31.03) | 174.36 (22.74) | 126.53 (39.27) |
| **TTR*** | Int. | .53 (.08) | .51 (.06) | .54 (.06) | .52 (.05) | .53 (.07) |
| | Ind. | .54 (.08) | .48 (.05) | .51 (.07) | .47 (.04) | .50 (.06) |
| **Total number of K1 tokens*** | Int. | 95.96 (33.05) | 119.16 (23.98) | 120.86 (21.82) | 143.00 (28.29) | 118.58 (30.57) |
| | Ind. | 66.71 (20.61) | 108.22 (25.68) | 112.89 (30.00) | 150.68 (24.20) | 107.62 (37.59) |
| **Total number of K2 tokens** | Int. | 5.12 (2.69) | 5.18 (2.82) | 5.75 (2.58) | 6.65 (3.81) | 5.61 (2.95) |
| | Ind. | 4.68 (2.82) | 4.94 (3.22) | 5.24 (2.65) | 6.05 (4.90) | 5.17 (3.36) |
| **Total number of AWL tokens** | Int. | .64 (.92) | .90 (.94) | 2.68 (2.66) | 2.49 (2.09) | 1.66 (2.03) |
| | Ind. | .64 (1.13) | 1.25 (1.24) | 1.79 (1.95) | 3.09 (1.57) | 1.61 (1.72) |
| **Lexical density*** | Int. | .52 (.06) | .52 (.06) | .52 (.04) | .51 (.05) | .52 (.05) |
| | Ind. | .49 (.06) | .48 (.05) | .46 (.05) | .44 (.04) | .47 (.05) |
| **Total number of word families*** | Int. | 55.48 (20.21) | 68.61 (17.35) | 68.15 (13.15) | 81.39 (18.09) | 67.73 (18.85) |
| | Ind. | 36.07 (8.35) | 49.81 (10.83) | 54.06 (10.30) | 66.77 (7.67) | 50.96 (13.93) |

*Note.* * *p*<.01 for overall analysis (lexical density was measured by the number of content words divided by total number of words).

** *Int*. represents interactive tasks. *Ind*. represents individual tasks.

*p* = .000; TTR, *t* (115) = 3.54, *p* = .001; total number of K1 (the most frequent 1,000 words of English) tokens, *t* (115) = 3.81, *p* = .000; lexical density, *t* (115) = 7.91, *p* = .000; and total number of word families, *t* (115) = 11.12, *p* = .000. The results indicate that more types, tokens, K1 tokens, K2 tokens, and word families were produced in the interactive task than in the individual task. Furthermore, examinees demonstrated a higher level of TTR and lexical density in the interactive task, compared to those in their individual task situation. Note that lexical density was measured by the number of content words divided by total number of words.

For the *Cambridge English: Preliminary* level, significant differences were found in the following variables: total number of types, *t* (27) = 5.31, *p* = .000; total number of tokens, *t* (27) = 4.94, *p* = .000; total number of K1 tokens, *t* (27) = 5.39, *p* = .000; lexical density, *t* (27) = 2.46, *p* = .021; and total number of word families, *t* (27) = 5.87, *p* = .000. This means that candidates at this level produced significantly more types, tokens, K1 tokens, and word families with a higher lexical density in the interactive task than in the individual task. In *Cambridge English: First*, six features were found to be significantly different between the two tasks: total number of types, *t* (31) = 4.21, *p* = .000; total number of tokens, *t* (31) = 2.89, *p* = .007; TTR, *t* (31) = 2.67, *p* = .012; total number of K1 tokens, *t* (31) = 2.35, *p* = .025; lexical density, *t* (31) = 3.51, *p* = .001; and total number of word families, *t* (31) = 6.20, *p* = .000. In the interactive task, examinees at this level produced significantly more types, tokens, K1 tokens, and word families with a higher lexical density than in the individual task. Meanwhile, they had a significantly higher level of TTR and lexical density in conversation. Four features showed significant differences at the C1 level. They were total number of types, *t* (33) = 4.52, *p* = .000; TTR, *t* (33) = 2.35, *p* = .025; lexical density, *t* (33) = 5.65, *p* = .000; and total number of word families, *t* (33) = 5.46, *p* = .000. Examinees at the *Cambridge English: Advanced* level seemed to have produced significantly more types and word families with a higher TTR and lexical density in the interactive tasks, compared to those in the individual task. Finally in *Cambridge English: Proficiency*, there were also significant differences found between the two tasks with the following four lexical variables: total number of tokens, *t* (21) = −2.09, *p* = .049; TTR, *t* (21) = 3.72, *p* = .001; lexical density, *t* (33) = 4.34, *p* = .000; and total number of word families, *t* (33) = 4.53, *p* = .000. At this level, the interactive task produced significantly fewer types but more word families, and yielded a higher level of TTR and lexical density than the individual task. Note that the patterns of token use were vastly different in this level, compared to the other three levels; i.e. candidates produced more tokens, and K1 tokens in the individual task.

### Intonation analysis results

As shown in Table 6 (see the Total column), three out of six features in intonation revealed statistically significant differences across four proficiency levels: level tone, *t* (115) =4.83, *p* = .000; space, *t* (115) = −4.39, *p* = .000; and pitch range, *t* (115) = 10.15, *p* = .000. It appears that examinees generally had more tone variation with a wider pitch range in the interactive task than in the individual task. They also demonstrated a lower number of prominent syllables per turn (pace) and a lower proportion of prominent words to the total number of words (space) in the interactive task.

At the B1 level, the following four variables showed statistical significance between the two tasks: level tone, *t* (27) = 3.01, *p* = .006; pace, *t* (27) = −6.36, *p* = .000; space, *t* (27) = −7.74, *p* = .000; and pitch range, *t* (27) = 5.12, *p* = .000. The results indicate that examinees in *Cambridge English: Preliminary* had more level tones and a wider range of pitch but with a lower level of pace and space in the interactive task. This means that candidates produced fewer numbers of prominent syllables by emphasizing content-related words in the interactive tasks, compared to their performances in the individual task. Five pronunciation variables were significantly different between the two tasks in *Cambridge English: First*: falling tone, *t* (31) = −2.46, *p* = .020; level tone, *t* (31) = 2.53, *p* = .017; pace, *t* (31) = −4.18, *p* = .000; space, *t* (31) = −5.87, *p* = .000; and pitch range, *t* (31) = 7.02, *p* = .000. That is, candidates started to show more intonation variation by using fewer falling and more level tones. In addition, they used a wider range of pitch and a lower level of pace and space in the interactive task. Significant difference was especially found in the following two variables in *Cambridge English: Advanced*: falling tone, *t* (33) = 4.34, *p* = .000; and pitch range, *t* (33) = 4.46, *p* = .000. Candidates at this level started to have significantly more falling tones in the interactive task than in the individual tasks. They also produced a wider range of pitch in the interactive task.

Three features between the two tasks turned out to be significant at the *Cambridge English: Proficiency* level: level tone, *t* (21) = 3.00, *p* = .007; space, *t* (21) =3.59, *p* = .002; and pitch range, *t* (21) = 3.82, *p* = .001. Candidates used more intonation variation with a wider range of pitch when having a conversation. The proportion of prominent words to the total number of words was lower in conversations than in monologues. That is, examinees tended to emphasize words more selectively (perhaps more meaning based) in the interactive task than in the individual task.

### Interaction analysis results

The descriptive statistics in Table 7 showed that higher-proficiency speakers tended to produce more co-operation moves such as back-channelling, prompting phrases and sentences, topic initiation and overall initiation. When it comes to discourse markers, higher-proficiency examinees generally had more discourse markers. As for turn-taking, higher-proficiency speakers had more turns between the two interlocutors. Short turns (fewer than 10 words) and middle turns (between 11 and 30 words) were used more frequently by higher-proficiency examinees. Longer turns (more than 30 words) were used more often among the lower-proficiency candidates. Concerning strategy use, higher-proficiency speakers tended to use more response maintenance strategies

**Table 6: Intonation features identified by proficiency levels**

| Features | Task | B1, Preliminary (N = 28) | B2, First (N = 32) | C1, Advanced (N = 34) | C2, Proficiency (N = 22) | Total (N = 116) |
|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Falling tone** | Int.** | 15.39 (5.65) | 13.03 (4.48) | 16.06 (6.17) | 19.91 (5.89) | 15.79 (5.97) |
| | Ind. | 14.64 (5.72) | 16.34 (7.50) | 10.44 (4.68) | 17.82 (6.51) | 14.48 (6.69) |
| **Level tone*** | Int. | 3.36 (2.97) | 4.44 (3.49) | 4.62 (3.62) | 5.23 (3.05) | 4.38 (3.35) |
| | Ind. | 1.25 (1.84) | 2.88 (2.32) | 3.44 (4.61) | 2.36 (3.47) | 2.55 (3.35) |
| **Rising tone** | Int. | 11.11 (3.77) | 11.63 (4.75) | 14.03 (4.11) | 14.36 (5.72) | 12.72 (4.72) |
| | Ind. | 10.68 (5.58) | 11.34 (4.25) | 13.12 (6.60) | 12.14 (6.17) | 11.85 (5.71) |
| **Pace** | Int. | .79 (.15) | .85 (.24) | 1.14 (.48) | 1.23 (.35) | .99 (.38) |
| | Ind. | 1.21 (.31) | 1.15 (.27) | 1.20 (.38) | 1.92 (3.75) | 1.33 (1.66) |
| **Space*** | Int. | .33 (.05) | .31 (.06) | .38 (.13) | .40 (.08) | .35 (.10) |
| | Ind. | .58 (.16) | .44 (.11) | .36 (.10) | .33 (.08) | .43 (.15) |
| **Pitch range*** | Int. | 118.63 (51.36) | 137.49 (49.59) | 137.93 (63.99) | 143.93 (47.29) | 134.29 (54.32) |
| | Ind. | 69.95 (40.46) | 73.23 (31.17) | 90.52 (44.57) | 98.39 (43.67) | 82.28 (41.16) |

*Note.* * *p* < .01 for overall analysis.

** *Int.* represents interactive tasks. *Ind.* represents individual tasks.

**Table 7: Interactive features identified by proficiency levels**

| Features | B1, Preliminary (N = 28) | | B2, First (N = 32) | | C1, Advanced (N = 34) | | C2, Proficiency (N = 22) | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **Back-channelling*** | 3.14 | 2.99 | 5.81 | 3.09 | 6.35 | 4.97 | 7.41 | 4.58 |
| **Prompting** | .86 | .80 | 1.13 | 1.07 | 1.18 | 1.00 | 1.41 | 1.44 |
| **Topic initiation*** | 2.04 | .96 | 3.09 | 1.49 | 5.00 | 2.26 | 3.59 | 1.62 |
| **Overlapping initiation** | 2.04 | 1.79 | 4.50 | 3.81 | 3.74 | 3.21 | 5.14 | 2.90 |
| **Discourse markers** | 8.43 | 3.72 | 8.16 | 3.43 | 9.26 | 4.41 | 10.68 | 4.06 |
| **Total talking time** | 56.36 | 13.30 | 61.78 | 10.99 | 58.56 | 11.03 | 59.30 | 18.38 |
| **Turn-taking time** | 9.55 | 4.47 | 10.52 | 4.84 | 9.53 | 3.08 | 11.78 | 4.96 |
| **Number of total turns*** | 7.29 | 4.38 | 15.41 | 6.33 | 14.35 | 7.10 | 17.86 | 5.33 |
| **Number of short turns*** | 3.96 | 3.87 | 10.88 | 6.35 | 9.82 | 6.51 | 12.41 | 5.50 |
| **Number of middle turns*** | 2.11 | 1.59 | 3.66 | 1.98 | 3.56 | 2.05 | 4.95 | 2.24 |
| **Number of long turns** | 1.21 | .79 | .88 | .71 | .97 | 1.00 | .50 | .80 |
| **Response maintenance*** | .68 | .77 | 1.28 | 1.25 | 1.74 | 1.42 | 2.82 | 1.65 |
| **Repair** | 3.04 | 2.67 | 2.19 | 1.69 | 2.32 | 2.00 | 2.05 | 2.10 |

*Note.* * *p* < .01 for overall analysis.

but fewer repair moves during the interaction, as compared to lower-proficiency candidates.

After a series of ANOVA tests, Tukey tests as post hoc analyses were conducted to identify significant group mean differences among four proficiency levels in terms of the frequency of occurrence on each interactive variable. Across all four levels, examinees demonstrated six statistically significant differences in interaction features: back-channelling (F3, 116=4.08; *p* =.009); topic initiation (F3, 116=15.98; *p* =.000), number of total turns (F3, 116=10.42; *p* =.000), number of short turns (F3, 116=8.22; *p* =.000), number of middle turns (F3, 116=5.37; *p* =.002), and response maintenance (F3, 116=6.18; *p* =.001).

Tukey tests revealed that there was a significant difference between *Cambridge English: Preliminary* and *Cambridge English: Proficiency* level candidates in using back-channelling actions. As for the topic initiation, significantly more topic initiations occurred as the proficiency increased. The increase was statistically significant between each adjacent level. The *Cambridge English: Preliminary* candidates had significantly fewer overlapping initiations than those in *Cambridge English: Proficiency*. A significant difference of using discourse markers was found between *Cambridge English: First* and *Cambridge English: Proficiency* levels. In terms of turn-taking time, the lowest proficiency level (*Cambridge English: Preliminary*) candidates had significantly fewer turns, but there was no simple linear pattern found across *Cambridge English: First, Cambridge English: Advanced* and *Cambridge English: Proficiency* levels. The use of short, middle, and long turns yielded a similar pattern: no simple linear pattern was found as the proficiency level increased. With regard to response maintenance, lower-proficiency candidates produced fewer actions of response maintenance compared to those of higher-proficiency candidates.

## Discussion and conclusion

The overall findings of the study suggest that there were distinctive differences in linguistic features in the two tasks across four CEFR speaking levels (*Cambridge English: Preliminary, Cambridge English: First, Cambridge English: Advanced* and *Cambridge English: Proficiency*). In other words, the task type impact on candidates' oral performance proves to be significant in L2 assessment. Accordingly, linguistic features extracted from an individual task, which may distinguish among proficiency levels, may not represent those from a paired/interactive task. Findings indicate a need for applying various task types and task-specific assessment scales in L2 speaking tests.

With regard to fluency variables in the criterion of discourse management, examinees demonstrated statistically significant differences generally in all four levels: in the interactive task, candidates produced more syllables per second, but shorter silent and filled pauses, compared to a testing context in the individual task. This means that examinees talked faster with a shorter duration of hesitation markers in conversation than in monologue. Examinees might sense the pressure of co-operating with a partner; accordingly, they might choose to increase their speed in conversation with a shorter period of complete silence. However, examinees in general had more filled pauses in the interactive task than in the individual long turn task. The increased use of filled pauses may represent the nature of interactive speech itself. That is, in the interactive task, examinees might have to collaborate and respond to their partner's conversation flow, while they had no such distraction when producing a monologue. Perhaps, filled pauses, unlike silent pauses, are a means of naturally holding the floor, which indicate an active participation in conversation, as it could indicate speakers' ongoing engagement. The results of separate paired-sample *t*-tests in each of the four levels revealed more insight into interpreting the fluency differences in the two tasks. Examinees in *Cambridge English: Preliminary* produced more syllables per second with more but shorter silent and filled pauses in the interactive task than the individual task. This finding is in line with the overall patterns. Given that examinees, especially at a lower level, might be less flexible in responding to their partner, the more frequent use of pauses in conversation is not too surprising. Examinees at the B2 level (i.e. *Cambridge English: First*) produced significantly fewer and shorter silent pauses but more filled pauses in the interactive task. This observation implies two possible interpretations. Firstly, similar to examinees at the B1 level, examinees that were slightly more proficient still appeared to produce more filled pauses instead of silent pauses when having a conversation. Secondly, unlike examinees at B1, examinees at this level used fewer silent pauses in conversation, which concurred with the overall observation across levels in which candidates used more filled pauses. As for the C1 and C2 levels, higher-proficiency examinees produced more syllables per second with shorter runs and pauses in the interactive task. This suggests that interaction was more actively taking place when examinees were more fluent. Largely, our findings indicated that candidates in a high stakes test environment tended to speak somewhat more fluently when they were involved in interaction.

Distinctive patterns were also found in the results of the grammatical complexity analysis. The results of all four levels combined showed that more T-units, error-free T-units, and clauses were produced in the interactive task than in the individual task. However, the overall T-unit complexity level was significantly higher in the individual task than in the interactive task, which showed that more dependent clauses were used in monologue than in conversation. In general, candidates tended to use more grammatically complex phrases and sentences in an individual task, whereas in an interactive task, their grammatical forms were simpler. A possible reason can relate to psychological stress perceived by examinees during an interactive task, where pressure may be present with interlocutors involved. The uncertainty of how a partner initiates and responds to the examinee's utterance could distract them from developing ideas in a more complex fashion.

As for the grammar complexity at each level, examinees in *Cambridge English: Preliminary* produced more T-units, error-free units, clauses, and dependent clauses with a significantly higher level of complexity in the interactive task than in the individual task. Examinees at higher proficiency levels such as *Cambridge English: First, Cambridge English: Advanced* and *Cambridge English: Proficiency*, although they used significantly more T-units and clauses when having a conversation, started to have a lower complexity level in conversation than in monologue. This observation suggests that with an interlocutor's presence, examinees across all levels had produced more utterances with a wider range of grammatical structures. However, those at higher proficiency levels showed a strong tendency to produce simpler grammatical structures in conversation as compared to those in monologic speech. This supports the importance of task differentiation in language assessment. In other words, candidates might cut back their lengthy expressions (e.g. dependent clauses) when interacting with others in the assessment context.

With the average of 17 types of grammatical errors, only six features were found to be significantly different between the tasks ($p < .01$). Examinees generally made more errors in using adjectives, verbs, subject and object in the interactive task than in the individual task. However, significantly fewer errors were made with regard to relative clauses and determiners in conversation. This is partially because in an interactive task, learners might not necessarily use a grammatical aspect as complex as the relative clause. At each level, no clear linear pattern was found in terms of total number of errors on certain types of grammar use. Examinees in *Cambridge English: Preliminary* had significantly more errors using prepositions, adjectives and objects in the interactive task than in the individual task. Examinees at the *Cambridge English: First* level produced significantly more errors in using verbs but fewer errors in using relative clauses in the interactive task than in the individual task. Examinees at *Cambridge English: Advanced* level had significantly more errors in using objects but fewer errors in tenses, determiners and relative clauses in the interactive task than in the individual task. Finally, learners at *Cambridge English: Proficiency* level had significantly more errors in using subject and copula in the interactive task than in the individual task. As no clear pattern emerged, our interpretation of these findings is somewhat limited at this stage. Additional studies might be needed to further explore the reasons of these findings.

Regarding lexical resources, the results of all four levels combined showed that more types, tokens, K1 tokens, K2 tokens and word families were produced in the interactive task than in the individual task. In addition, examinees demonstrated a higher level of TTR and lexical density in conversation. The results of these differences in lexical resources between the two tasks indicated that the interactive task elicited a wider range of lexical features with a higher TTR and lexical density than the individual task. Examinees performed better in lexical richness when they interacted with others. The results at each level further illustrated this pattern: in the interactive task, the examinees at B1, B2 and C1 levels generally produced significantly more types, tokens, K1 tokens and word families with a higher lexical density. The examinees overall had a higher TTR in the individual task than that in the interactive task. This result can be a function of the greater range of visual prompts in the interactive task, which may provide opportunities to discuss on a wider range of topics and hence a richer performance in lexical resources. One exception was the use of tokens among the most proficient examinees at the C2 level, where they used more tokens in the individual task.

Our results from the intonation analysis were interesting. The examinees generally had significantly more tone variation with a wider pitch range in the interactive task than in the individual task, which is somewhat expected, due to the nature of communication involvement. Intonation is critical in successful communication (Kang et al 2010, Pickering 2001). In a discourse context, speakers tend to alter the dynamic of their pitch or tone to successfully deliver their intended message. As a way to signal their participation in the conversation, interlocutors may constantly try to reach the agreement of their pitch level (Kang 2010) or change their pitch or tone to hold the conversational floor. As a result, in order to achieve communication goals effectively, examinees used more varying intonation patterns in an interactive task than in a monologic individual task. The examinees also demonstrated a lower number of prominent syllables per turn (pace) and a lower proportion of prominent words to the total number of words (space) in the interactive task than in an individual task. Low-proficiency speakers tended to place stress on words (regardless of their functions) more frequently in monologues than in conversations (Kang 2013). Typically, low-proficiency non-native speakers in the monologic speech use primary stress on every lexical item, regardless of its function or semantic importance (Kang 2010, Wennerstrom 2000). Nevertheless, the candidates might have altered this pattern by accommodating their needs to interact with their partners successfully. The wider range of pitch implies that examinees performed more naturally when having a conversation with their interlocutors than when giving a monologue in the context of oral assessment.

Some informative patterns also emerged from the interaction analysis on the interactive task. The descriptive statistics showed that the higher-proficiency level speakers tended to produce more co-operation moves such as back-channelling, prompting phrases and sentences, topic initiation and overall initiation. The higher-proficiency examinees also generally used more discourse markers. The results suggested that the higher-level candidates were more active in using a variety of interaction features to demonstrate their ability to co-operate with a partner in conversations. Moves such as back-channelling, prompting, and initiating a new topic suggest an improved degree of engagement in a conversation. The higher-proficiency speakers had more turns between each interlocutor. Short turns (fewer than 10 words) and middle turns (between 11 and 30 words) were used more frequently by the higher-proficiency examinees (Galaczi 2013). Longer turns (more than 30 words) were used more among the lower-level candidates. These results indicate that the candidates had more interactions and were involved more in each other's ideas by exchanging turns more promptly. The greater use of short and middle turns helped speed up the turn-taking process between interlocutors. Conversely, the lower-proficiency level examinees used fewer turns with each other since they tended to produce longer turns, which led to fewer turn-switches. When it comes to strategy use, it seemed that the higher-proficiency speakers used more response maintenance strategies but fewer repair moves during the interaction. More actions on response maintenance such as repeating words or phrases from their partner could help build rapport and create a smoother conversation flow between the two interlocutors. It is not surprising to see more use of this strategy among the higher-proficiency examinees. The lower-proficiency candidates, on the other hand, had more repair moves than the advanced candidates. As a result, a decrease of repairing one's own utterance seemed to indicate a higher proficiency level.

To conclude, the study demonstrated task type differences in a wide range of linguistic features including fluency, grammatical resources, lexical resources, and pronunciation. Overall, compared to an individual task, examinees across all levels seemed to talk faster with shorter periods of pauses in an interactive task. The examinees in general also demonstrated a wider range of grammatical and lexical resources at a higher complexity level. Finally, the examinees had a less restricted pitch range and a more natural intonation performance when having a conversation instead of a monologue. The results of the current study imply that compared to a more traditional individual task, an interactive task can offer opportunities to elicit more diverse and natural linguistic performance in the context of oral assessment. The results serve as empirical evidence for integrating conversations into both classroom instruction and assessment on L2 speaking skills.

At the same time, our findings confirm that task types do make a significant difference in assessing speaking ability. This difference should be carefully considered in several contexts. Especially in the context of English speaking assessment, the difference between the tasks should be accounted for and integrated in developing scoring scales and rater training. In fact, the majority of Cambridge English Speaking tests do involve both monologic and interactive tasks to address differences across task types. The current findings provide validity support for the Cambridge English Speaking test scores and assessment scales. The Speaking scores used in the analysed data are based on the individual judgement of Cambridge examiners and on empirically based assessment scales. This study, coming from a different angle, has used objective measures and has corroborated differences in proficiency levels.

Finally, the interaction analysis in this study revealed some significant relationships between interaction features and proficiency levels. The more advanced examinees demonstrated a more active engagement of conversations by using a variety of features like back-channelling, prompting, initiating a new topic, more interactive turn-taking management, and conversation strategies, compared to the less advanced examinees. The findings have at least two important implications. Firstly, in the L2 speaking classroom, more explicit instruction on these features is needed to help improve examinees' interaction skills. Secondly, introducing these features to future scoring systems for conversations may be important as it will better capture candidates' communication performance. Although more task-specific scales may require a higher level of cognitive loads for both examiners and raters (Taylor and Galaczi 2011), assessment scales and examiner training materials can benefit from explicit descriptors of certain linguistic features which are found to play a distinguishing role across proficiency levels.

## References

Anderson-Hsieh, J, Johnson, R and Koehler, K (1992) The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure, *Language Learning* 42, 529–555.

Brown, G and Yule, G (1983) *Discourse Analysis*, Cambridge: Cambridge University Press.

Brown, A, Iwashita, N and McNamara, T (2005) *An Examination of Rater Orientations and Test-taker Performance on English-for-Academic-Purposes Speaking Tasks*, TOEFL Monograph Series MS-29, Princeton: Educational Testing Service.

Celce-Murcia, M, Dörnyei, Z and Thurrell, S (1995). Communicative competence: a pedagogically motivated model with content specifications, *Issues in Applied Linguistics* 6, 5–35.

Crookes, G (1990) The utterance, and other basic units for second language discourse analysis, *Applied Linguistics* 11, 183–199.

Galaczi, E D (2013) Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, available online: doi:10.1093/applin/amt017.

Gnisci, A and Bakeman, R (2007) Sequential accommodation of turn taking and turn length: A study of courtroom interaction, *Journal of Language and Social Psychology* 26 (3), 234–259.

Hunt, K W (1970) Syntactic maturity in school children and adults, *Monographs of the Society for Research in Child Development* 35 (1), 1–61.

Iwashita, N, Brown, A, McNamara, T and O'Hagan, S (2008) Assessed levels of second language speaking proficiency: How difficult? *Applied Linguistics* 29, 24–49.

Kang, O (2008) Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness, *Spaan Fellow Working Papers* 6, 181–205.

Kang, O (2010) Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness, *System* 38 (2), 301–315.

Kang, O (2013) Linguistic analysis of speaking features distinguishing general English exams at CEFR levels B1 to C2 and examinee L1 backgrounds, *Research Notes* 52, 40–48.

Kang, O and Wang, L (2013) *Linguistic features distinguishing examinees' speaking performance at different proficiency levels*, paper presented at the American Association of Applied Linguistics (AAAL) 2013, Dallas.

Kang, O, Rubin, D and Pickering, L (2010) Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English, *Modern Language Journal* 94, 554–566.

Kim, Y (2009) The effects of task complexity on learner-learner interaction, *System* 37, 254–268.

Kormos, J and Denes, M (2004) Exploring measures and perceptions of fluency in the speech of second language learners, *System* 32, 145–164.

Nakatani, Y (2010) Identifying strategies that facilitate EFL Examinees' oral communication: A classroom study using multiple data collection procedures, *The Modern Language Journal* 94, 116–136.

O'Sullivan, B, Weir, C and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19 (1), 33–56.

Pickering, L (2001) The role of tone choice in improving ITA communication in the classroom, *TESOL Quarterly* 35, 233–255.

Riggenbach, H (1991) Towards an understanding of fluency: A microanalysis of nonnative speaker conversation, *Discourse Processes* 14, 423–441.

Schegloff, E A (1982) Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences, in Tannen, D (Ed) *Analyzing Discourse: Text and Talk*, Washington, D.C.: Georgetown University Press, 71–93.

Skehan, P (2001) Tasks and language performance assessment, in Bygate, M, Skehan, P and Swain, M (Eds) *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*, London: Longman, 210–228.

Taylor, L and Galaczi, E D (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 171–233.

Wennerstrom, A (2000) The role of intonation in second language fluency, in Riggenbach, H (Ed) *Perspectives on Fluency*, Ann Arbor: University of Michigan Press, 102–127.

# Examiner confidence survey: An investigation into Speaking Examiners' confidence in the accuracy of the assessments they make

**SUE GILBERT** CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT, SWITZERLAND
**GEORGIA STAUB** CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT, SWITZERLAND

## Introduction

This study was undertaken in response to an occasional feeling of uncertainty when making assessments of speaking skills which was reported by some Speaking Examiners (SEs) in Switzerland in relation to the *Cambridge English: Advanced* Speaking test. The SEs involved were all experienced and reliable raters, and so it was decided that the issue merited investigation since it could provide insights about the training needs of the examiners involved, and could also contribute to a more in-depth understanding of the rating process of trained examiners in general. Speaking assessment is a complex process since assessment decisions are based on a wide range of interacting factors which need to be taken into consideration in real time when deciding on a mark. It is important, therefore, to explore periodically the issues which examiners face during their decision-making process and to focus on areas of difficulty which they encounter. SEs were asked to rate and were invited to comment on their confidence in relation to aspects of assessment. The objective was to see whether useful insights could be gained which might feed into any aspect of the Speaking test design or SE training and professional development, and so lead to improved certainty in arriving at assessments.

## Context of the study

Research studies have focused on candidates' experience of Speaking tests (for example Chambers, Galaczi and Gilbert 2012, Humphry-Baker 2000); this study investigates SEs' perceptions related to how they assess speaking. Specifically, it is an attempt to gain insight into how SEs perceive the reliability of their own assessments and the factors which may impact on the marking process. We must consider the possibility that building SEs' confidence in the reliability of their own rating skills could perhaps be a factor in increasing reliability itself.

It is useful in the context of this study to understand the duties of a Cambridge English Language Assessment SE. Cambridge English Language Assessment SEs are required to fulfil two different assessment roles during Speaking tests: to act as Interlocutor and Assessor. The Interlocutor manages the tests, interacts with the candidates and awards a holistic assessment for each candidate based on a Global Achievement scale. The Assessor does not interact with the candidates but refers to three to five different scales, depending on the test, to make an analytical assessment of each candidate's performance. SEs are required to perform both roles during each examining assignment. As part of ongoing Quality Assurance all Cambridge English SEs must successfully complete an annual certification process, including both online and face-to-face components. Both Quality Assurance and professional development of SEs are managed by Team Leaders (TLs), who run the face-to-face element of certification in the form of an obligatory meeting which includes a focus on professional development. TLs use their monitoring visits to Speaking tests, feedback from Centre Exams Managers, development requests from SEs in their teams, statistical monitoring carried out in Cambridge, and their own experience as an SE working in the team to establish the focus of professional development activities, based on perceived need. Then, with the support of the Professional Support Leader for their country or region, they select or design activities aimed at supporting SEs in improving their examining skills in the area identified. This study forms part of the process of investigating a possible perceived need for SEs' specific professional development, on an issue raised by SEs themselves.

The issue of SEs' confidence in the accuracy of the marks they award could have arisen as a result of their application of reflective practice, which has long been known to be an important element in successful professional development in many professions. The value of thinking over and questioning one's own actions and decisions is generally seen as extremely helpful in improving performance. For example, the Ofsted 2004 report *Why Colleges Succeed* (2004:10) states: 'The most distinctive characteristic of these very good teachers is that their practice is the result of careful reflection,[. . .] they themselves also learn lessons each time they teach, evaluating what they do and using these self-critical evaluations to adjust what they do next time.'

In her article *The Reflective Teacher*, Moon (2005:15) sums up the value of reflection: 'Reflective practice is the key to improvement. If we don't think about, analyse and evaluate our professional practice we cannot improve.' Similarly, in their article *The Making of an Expert*, Anders Ericsson, Prietula and Cokely refer to the importance for experienced professionals of maintaining the ability to analyse a situation and work through the right response, and of 'deliberate practice' in the process of developing expertise (2007:4–5).

Teaching experience is an essential criterion in the recruitment of SEs. It is therefore logical to assume that the habitual process of reflection and analysis will also occur spontaneously in the case of examining, even though the requirements of the role are very different from that of teaching. In fact the importance of supported personal reflection as part of the process of improving expertise among Cambridge English SEs could be argued to have increased in recent years, as a considerable part of the certification process is now carried out online, and therefore without immediate input from a Team Leader. It is therefore important that SEs feel they have the resources and skills to manage their own professional development. The value of reflection in professional development is recognised in the 2013 *Guidelines for Speaking Examiner Certification*, which state (Cambridge English Language Assessment 2013:8) that 'The aim [of annual certification] is to support examiners in their approach to the assessment of Speaking tests. Examiners need to be encouraged to reflect on how they assess and, if necessary, to make adjustments that are meaningful and permanent.'

When three TLs from different examination centres in Switzerland, and a majority of the examiners in one of the three centres reported that they experienced occasional difficulty in finalising one or more of the marks while assessing some *Cambridge English: Advanced* candidates, it was understood in the context of their habitual professional reflection, meaning that these are competent SEs searching to improve their own performance, rather than SEs who are not performing satisfactorily. In fact, all the SEs involved in raising the question have consistently produced satisfactory ratings in the certification and monitoring processes, in some cases for more than 15 years, so there is no evidence that their assessments are unreliable. The SEs concerned reported that difficulty in finalising some marks had led to an occasional feeling of uncertainty when assessing, which caused them some unease, and to reflect on why this might be happening.

SEs' confidence in their accuracy in applying the assessment scales and awarding marks has been the subject of past investigations. In their worldwide survey of the views and experience of SEs running the *IELTS* Speaking test, Brown and Taylor (2006) found a reassuringly high (71%–91%) level of agreement with the affirmation 'I feel confident that my ratings are accurate when applying the scales'. Their further investigation into how confident SEs felt in applying the individual assessment scales showed that by far the largest proportion of examiners selected Pronunciation as the scale they felt least confident in applying. The two final questions in their survey asked SEs which aspects of the Speaking test they felt least and most confident about: '. . . many examiners were most confident about conducting the interview and following the script, whereas many were least confident about making accurate assessments' (2006:17). This seems to reflect the reported occasional lack of confidence in assessments which led to this study.

Yates, Zielinski and Pryor (2008) investigated how confident SEs felt in applying the different aspects of the *IELTS* Pronunciation scale, and concluded that 'while the examiners felt confident judging all of the features covered in the Pronunciation scale descriptors, they were more confident in judging the global features [ . . . ] than in judging most of the concrete features'. The SEs in Switzerland who reported their occasional lack of confidence did not refer specifically to the scales, but to an occasional general sense of unease about awarding marks, so a wider investigation around the various aspects of assessment seemed appropriate.

The topic was discussed at the annual Swiss TL meeting, and other TLs in the group supported the view that this was an area which should be investigated, with the aim of gaining useful insights into the extent and nature of any perceived insecurity in rating, so that any issues arising could be addressed as a part of professional development. Specifically, it would be useful to know whether examiners in other Swiss centres also experienced occasional difficulty when rating candidates, and whether the feeling applied only to Cambridge English: Advanced *(CAE)*, or to other Speaking tests/Common European Framework of Reference (CEFR) (Council of Europe 2001) levels as well.

In addition, the study aimed to see if it was possible to determine any factors which SEs perceived as impacting on their accuracy and reliability in the rating process. The hope was that increased insight into SEs' perceptions of the various factors involved in making reliable assessments might offer valuable information for SE trainers and Speaking test designers. In this way the study could prove of benefit to a wider audience. It was decided to focus on the following questions:

1. How confident do SEs in Switzerland feel about the reliability of their marking in the two different roles of Interlocutor and Assessor, and across the range of Speaking tests?
2. Can factors be identified which appear to affect rater confidence?

## Methodology

The investigation consisted of two stages, and two instruments were used:

Stage 1 consisted of a selected-response questionnaire, which included the option to add comments relating to each question. It was divided into two sections, focusing respectively on Global and Analytical rating (see Appendix 1, survey questions). The first aim was to investigate the level of confidence SEs have in the reliability of both their Global and Analytical rating of candidate performance in the range of General and Business English Cambridge Speaking tests. In questions 5–8 of the survey, participants were asked to rate their confidence (from 'very confident' to 'not confident') across tests when awarding Global Marks and Analytical Marks.

The second aim was to explore the possible factors which SEs may perceive as affecting their confidence in rating candidate performance. They were asked to indicate which, if any, of the following factors they felt could limit their confidence:

- having both to manage procedure and award marks (Global Marking only)
- relating performance to descriptors
- absence of expanded descriptors for some bands
- lack of training
- lack of experience.

Multiple answers were possible. Throughout the survey, it was not obligatory to answer any of the questions, and SEs had the option of writing additional extended answers to the questions.

Stage 2 consisted of individual and group interviews following up on the questions in the survey, with a cross-section of examiners from different Cambridge exam centres. The aim of the interviews was to explore further the areas under investigation. The questionnaire was used as the basis for the interviews, with additional follow-up questions where appropriate.

## Profile of Speaking Examiners involved in the study

112 SEs out of a total of approximately 160 (70%) from all the Cambridge English examination centres in Switzerland participated in the voluntary survey in August of 2012. Between 95 and 112 SEs answered each of the questions (it was permitted to leave questions unanswered).

To give a more detailed profile of the SEs involved, of the 100 respondents to the personal information questions, 89% are female, 87% speak English as their native language and 4% are bilingual in English and another language. Of responding SEs, 8% are aged 31–40, 35% are aged 41–50, 25% are aged 51–60 and 32% are 61 or over. As is shown in Figure 1, the 100 responding SEs teach at the following levels on the CEFR (Council of Europe 2001) in descending order: B2 (95%), C1 (88%), B1 (83%), A2 (56%), C2 (46%) and A1 (35%).

All 112 SEs responded to the question 'How long have you been a Cambridge English Speaking Examiner?'. As shown in Figure 2, 33 SEs (29.5%) have been examining for Cambridge English for longer than 15 years, 13 SEs (11.6%) for 11–15 years, 27 (24.1%) for 6–10 years, 16 (14.3%) for 2–5 years and 23 (20.5%) for two years or less.

111 SEs gave information about which exams they are certified for. Figure 3 shows the number of SEs qualified for each exam, indicating that most SEs are qualified for multiple exams.

The participating SEs therefore represent both male and female SEs, who are native and non-native English speakers and with a broad range of experience, from about a fifth who are relatively inexperienced (two years or less), to almost a third with more than 15 years' examining. The teaching background of the SEs (in relation to the CEFR) broadly reflects the relative candidature for Cambridge exams in Switzerland, and the exam qualifications information shows that almost all SEs are qualified for more than one exam, meaning they are able to make comparisons of their experience in assessing different tests.

**Figure 1: CEFR levels taught by SEs**

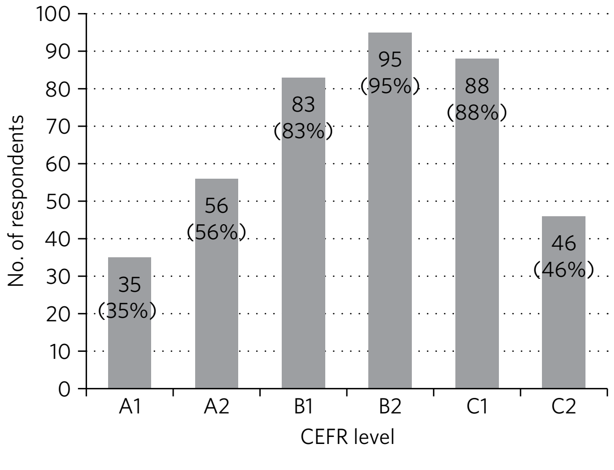


**Figure 2: Number of years participants have been Cambridge English Speaking Examiners (112 responding SEs)**

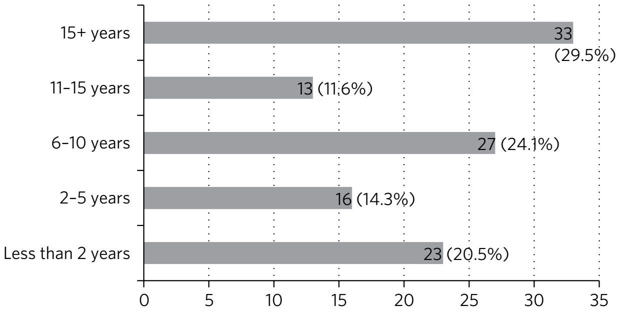
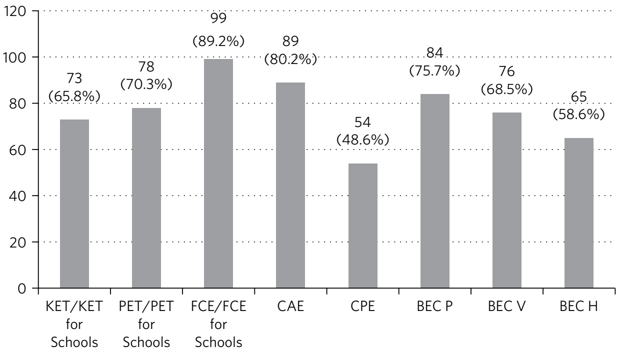


**Figure 3: Exam qualifications of 111 responding SEs**



Background information on SEs is not routinely collected, but from our deeper knowledge of some individual teams in Switzerland, we can say that the profiles reflect that of the Swiss cadre of SEs.

## Results and discussion

Results and discussion are organised as follows:

*Examiners' confidence*. Findings are reported relating to levels of confidence in making global and analytical assessments in different exams and when applying the different marking scales.

*Factors affecting confidence*. Findings are reported on SEs' opinions on a range of other factors which may influence their confidence when giving global and analytical assessments across the range of exams.

*Teaching experience and confidence*. SEs' views are reported on the importance of teaching experience in making assessments.

*The perceived helpfulness of various factors in building confidence*. Examiners' views on how they can build confidence are reported.

### Examiners' confidence

*Examiners' confidence in the two roles*

111 SEs responded to the question about their confidence in the two roles they undertake. Eighty-five (76.6%) of the respondents indicated that they feel equally confident as Interlocutor or Assessor. Nineteen (16.8%) felt more confident as Interlocutor, and seven (6.5%) as Assessor. These results might suggest that the examining role in itself may not be of importance in the question of rater confidence, and therefore it is necessary to investigate specific factors linked to the experience that examiners bring to the role, or to specific aspects of examining and assessment procedures or materials.

**Table 1: Examiner confidence when awarding Global Marks**

| | Very confident | Confident | Occasionally unconfident | Not confident | N/A | Response count |
|---|---|---|---|---|---|---|
| **KET** | 28.8% (21) | 42.5% (31) | 17.8% (13) | 0.0% (0) | 11.0% (8) | 73 |
| **PET** | 34.2% (26) | 43.4% (33) | 15.8% (12) | 0.0% (0) | 6.6% (5) | 76 |
| **FCE** | 43.5% (40) | 43.5% (40) | 13.0% (12) | 0.0% (0) | 0.0% (0) | 92 |
| **CAE** | 30.1% (25) | 45.8% (38) | 22.9% (19) | 0.0% (0) | 1.2% (1) | 83 |
| **CPE** | 20.4% (11) | 38.9% (21) | 29.6% (16) | 0.0% (0) | 11.1% (6) | 54 |
| **BEC P** | 34.1% (28) | 32.9% (27) | 25.6% (21) | 3.7% (3) | 3.7% (3) | 82 |
| **BEC V** | 36.5% (27) | 37.8% (28) | 18.9% (14) | 2.7% (2) | 4.1% (3) | 74 |
| **BEC H** | 26.2% (17) | 46.2% (30) | 20.0% (13) | 0.0% (0) | 7.7% (5) | 65 |

**Table 2: Examiner confidence when awarding Analytical Marks**

| | Very confident | Confident | Occasionally unconfident | Not confident | N/A | Response count |
|---|---|---|---|---|---|---|
| **KET** | 25.7% (18) | 44.3% (31) | 20.0% (14) | 0.0% (0) | 10.0% (7) | 70 |
| **PET** | 34.7% (25) | 47.2% (34) | 12.5% (9) | 0.0% (0) | 5.6% (4) | 72 |
| **FCE** | 36.3% (33) | 53.8% (49) | 9.9% (9) | 0.0% (0) | 0.0% (0) | 91 |
| **CAE** | 22.2% (18) | 49.4% (40) | 27.2% (22) | 0.0% (0) | 1.2% (1) | 81 |
| **CPE** | 16.7% (9) | 44.4% (24) | 25.9% (14) | 0.0% (0) | 13.0% (7) | 54 |
| **BEC P** | 28.8% (23) | 45.0% (36) | 21.3% (17) | 0.0% (0) | 5.0% (4) | 80 |
| **BEC V** | 28.8% (21) | 50.7% (37) | 15.1% (11) | 0.0% (0) | 5.5% (4) | 73 |
| **BEC H** | 23.1% (15) | 47.7% (31) | 18.5% (12) | 0.0% (0) | 10.8 % (7) | 65 |

*Levels of confidence in making Global and Analytical Assessments*

From the information reported in Table 1, we can see that a clear majority of between 59.3% (*Cambridge English: Proficiency*, also known as *Certificate of Proficiency in English (CPE))* and 87% (*Cambridge English: First*, also known as *First Certificate in English (FCE))* of the 105 SEs who responded to the question 'As Interlocutor, how confident do you feel awarding Global marks in each exam?' indicated that they felt 'confident' or 'very confident' when awarding Global Marks to any exam. 'Occasionally unconfident' was indicated by 29.6% when examining CPE, followed by 25.6% examining *Cambridge English: Business Preliminary (BEC P)* and by 22.9% examining *Cambridge English: Advanced*, also known as *Certificate in Advanced English (CAE)*. 'Not confident' in Global Assessments was selected only five times, three for *BEC P* and two for *Cambridge English: Business Vantage (BEC V)*.

A total of 102 SEs responded to the question 'How confident do you feel when awarding Analytical marks in each exam?' Table 2 shows that a clear majority of between 61% *(CPE)* and 90% *(FCE)* of respondents reported feeling 'very confident' or 'confident' when awarding Analytical Marks to any exam. 27% of the respondents felt 'occasionally unconfident' when assessing *CAE* followed by 26% examining *CPE* and 21% examining *BEC P*. No examiner described themselves as 'not confident' on any of the exams.

Thus we can see that overall levels of confidence are reassuringly high, but SEs identify the same exams – *BEC P*, *CAE* and *CPE* – most often as those where they are relatively less confident when making either Global or Analytical Assessments.

There are three analytical scales for the A2 test *(KET)*; Grammar and Vocabulary, Pronunciation, and Interactive Communication. B1 and B2 tests (*PET*, *FCE*, *BEC P* and *BEC V*) have an additional scale for Discourse Management and there are five scales for C1 and C2 tests (*CAE*, *BEC H* and *CPE*), where the Grammar and Vocabulary scale is separated into two individual scales: Grammatical Resources and Lexical Resources.

Respondents were asked to rate their confidence (from 'very confident' to 'not confident') in applying each of the analytical scales across the range of Speaking tests (see Figure 4). When assessing Interactive Communication, 87.2% of responding SEs expressed they were 'confident' or 'very confident', as compared to 85.2% on Grammar, 83.2% on Vocabulary, 81.1% on Pronunciation and 72.5% on Discourse Management. No SEs reported feeling 'not confident' on any of the scales: however, more than a quarter (27.5%) of SEs reported feeling 'occasionally unconfident' when assessing Discourse Management (DM).

Some SEs commented on their difficulty with the DM scale:

**Figure 4: Responses to Survey Question 9: As Assessor, how confident do you feel applying each of the analytical criteria?**

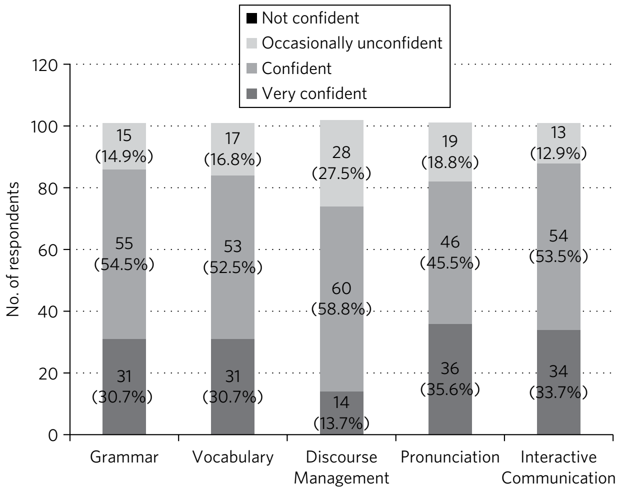
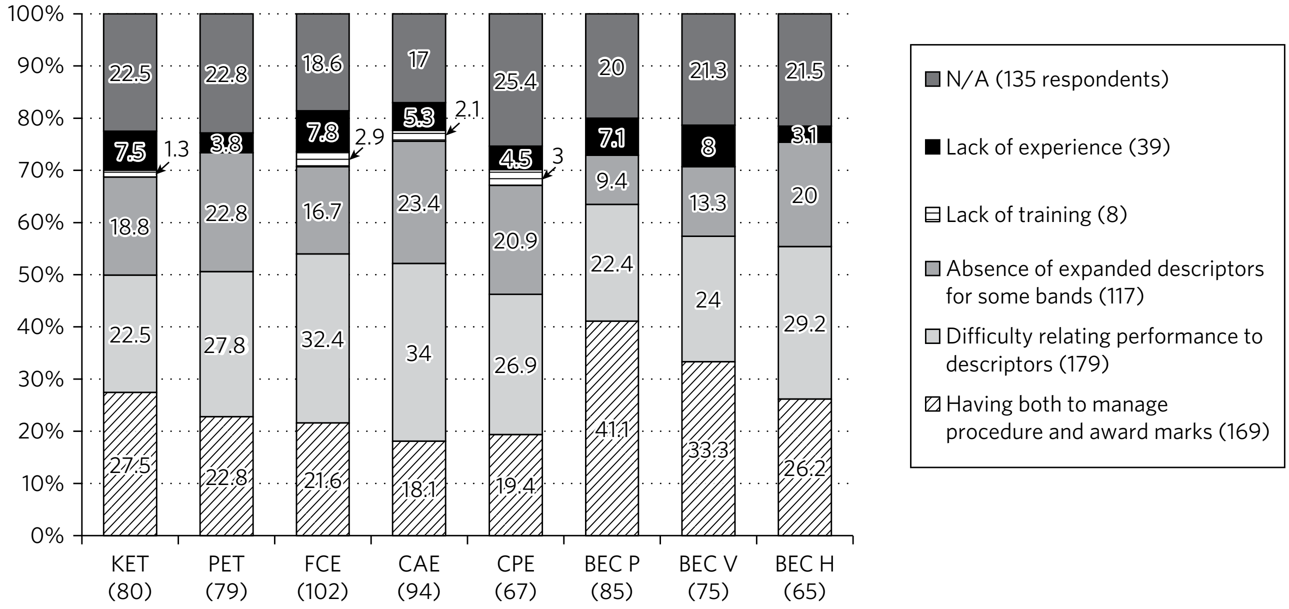


**Figure 5: Responses to Survey Question 6: As Interlocutor, what factors can limit your confidence when awarding Global Marks?**



*I would like to be clearer about how to interpret the IC and DM scales across the CEFR levels.*

*With DM I often have a problem with 'uses a range of cohesive devices'. Very often candidates don't use cohesive devices but still have good DM.*

*The descriptors for DM are useful in identifying a level higher than FCE [i.e. CEFR C1 and C2]. The idea of Discourse Markers is introduced here and that gives me something to hold on to.*

*DM can be difficult to assess, especially at lower levels when the teacher has taught the students a lot of connecting words, for example. I never know if they're using them or just parroting what they've learned by heart. It's almost as if they have connectors, but no language to connect.*

It could be useful for future SE training programmes to investigate this further, with a view to providing more focused training in interpreting the DM scale.

*Factors which may limit confidence in making assessments*

SEs could select as many or as few of the factors listed which might affect their confidence in making assessments (see Figures 5 and 6), and were invited to elaborate or give any other causes in written comments. As the Figures show, taking both Global and Analytical Marking, three factors were most often selected: 'difficulty relating performance to descriptors' was selected 357 times across all exams, 'not applicable (N/A)' 311 times and 'absence of expanded descriptors for some bands' was selected 280 times. In the case of Global Marking, 'having both to manage procedure and award marks' was selected 169 times. 'Difficulty relating performance to descriptors' was indicated 179 times in Global Marking and 178 in Analytical Marking. In their comments, SEs reflected on why they may have difficulty relating performance to the descriptors:

*Sometimes it can be difficult to relate a specific candidate's performance to the [Global Achievement] descriptors if the candidate is all over the place (e.g. good range of vocabulary but poor control of grammar, good interactive communication but poor pronunciation).*

*KET can be difficult sometimes because there is not a lot of language to assess. There are times when the test is over in 6 minutes and the candidates have said very little – apart from yes, no or two word answers.*

*The descriptors for the Global Mark are so vague that you have to have the Common Scales in front of you to understand the level.*

'Having both to manage procedure and award marks', which applies only to Global Marking, was selected 169 times. Comments from SEs illustrate their concerns:

*Sometimes difficult to award Global Marks as interlocutor needs to concentrate on so many different things.*

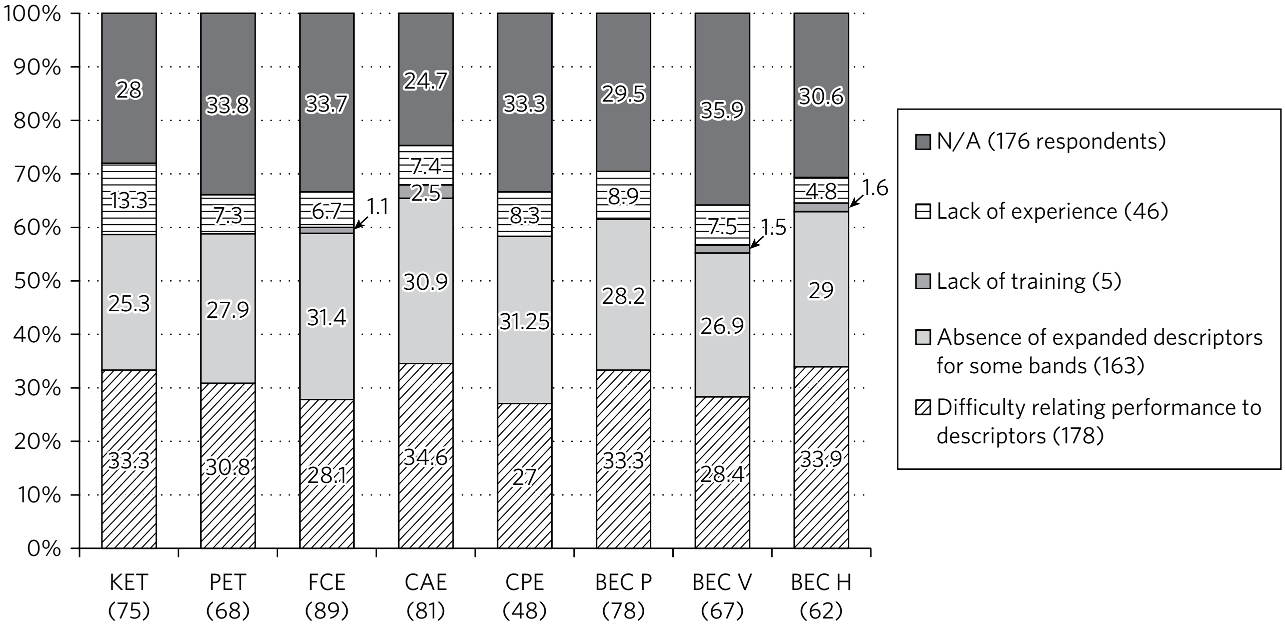
*I have to be careful in BEC V and BEC H not to get so caught up in delivering the exam that I don't at several moments in the test think about the Global Mark.*

*Sometimes I feel quite distracted by the chore of keeping to the rubrics and handling the timing in the BEC exams so I am less able to give a confident Global Mark.*

*It comes down to familiarity. KET I do find tricky because it's short . . . it's the shortest. If there are difficulties and you have to direct the candidates . . . prompt or point, that means your concentration is elsewhere.*

'N/A (Not applicable)' was selected 135 times in Global Marking and 176 times in Analytical Marking. We must allow for the fact that this figure includes SEs who are not qualified for the exam in question, but it could nevertheless suggest that there is an encouragingly high general level of confidence among SEs, The responses could also suggest that SEs are more confident when making Analytical Assessments, which allow them to focus on a single task, than making Global

**Figure 6: Responses to Survey Question 8: As Assessor, what factors can limit your confidence when awarding Analytical Marks?**



Assessments, when they have the dual role of conducting the test and awarding a mark.

'Absence of expanded descriptors for some bands' was also considered to be a factor by some SEs (selected 117 times across exams for Global Assessment and 163 times for Analytical Assessment).

*Some candidates just fall between bands and sometimes a lack of descriptors for the in-between marks makes it difficult to go up or down.*

This seems to contradict research which suggests that assessment scales have to be brief to be operationally effective (Council of Europe 2001:205–207).

'Lack of experience', selected 39 times in Global Marking and 46 in Analytical Marking, was described by some SEs as having an impact:

*The exams where I am occasionally unconfident are the exams I do not examine regularly.*

*I think it would be helpful to have more dummy runs as Interlocutor in training and co-ordination meetings, which concentrate on assessments.*

*CPE I'd like more practice – I do find that quite difficult. I find there's adequate time to listen to them, it's quite easy to administer, 2 minutes . . . So it's not a problem having enough of a sample, I think I need more practice or perhaps more exposure to Public Services Network (PSN) as well. It comes down to a lack of practice.*

It is also worth noting here that in both examining roles one factor affecting confidence was selected much less frequently: lack of training was selected eight times for Global Marking and five for Analytical.

*Factors affecting examiner confidence in different exams, Global and Analytical Marking*

Figures 5 and 6 show the number of times the various Speaking tests were selected for each factor. The single factor selected most frequently (by 41.1% of SEs) was 'having both to manage procedure and award marks' for BEC P, and for this factor all the BEC tests were cited more often than the equivalent level core tests. Comparing the two assessment roles, SEs report more difficulty relating performance to descriptors when awarding analytical marks than when marking globally. The absence of expanded descriptors was noted more in analytical marking than in Global Marking. In both roles 'difficulty relating performance to descriptors' was selected for *CAE* slightly more frequently than for other tests. As reported above, lack of experience and lack of training appear to impact less than other factors on examiner confidence across all the tests, which suggests that SEs are satisfied with their training and examining opportunities. However, 13.3% of SEs selected 'lack of experience' as a factor when making analytical assessment in *KET*, considerably more than other tests.

Multiple answers were possible. N/A could be selected where SEs are not qualified for the exam in question, or where they felt entirely confident in their marking. The total number of responses is noted in brackets for each test and for each factor.

SEs' comments on Analytical Marking seemed to focus on two aspects of applying the descriptors; interpreting the criteria in live tests and the interpretation of standardisation samples.

*Comments on applying the descriptors: Live tests*

In considering the practical challenges of applying the descriptors during live examinations, SEs indicated a range of issues in their comments. Their focus tended to be on the need to understand better the application of some aspects of the descriptors, sometimes in relation to particular exam levels. The design of particular tests in relation to the language sample produced for assessment was also commented on.

*CAE and BEC P are the two where I sometimes feel a need for more descriptors.*

*The differences in the descriptors are sometimes very fine when one considers the range of marks they cover, e.g. some/very little repetition (either a 1 or 3 for Discourse Management at FCE).*

*I have more difficulty in relating performance to descriptors at CAE than other MS [Main Suite] tests.*

*BEC P/V/H: not enough candidate input (too much interlocutor talking time).*

*Words like 'hesitation' as opposed to 'some hesitation' or 'mostly intelligible' as opposed to 'intelligible' can lead to feelings of uncertainty.*

*I think CAE and CPE are harder to examine and one reason might be a shifting relationship between the lexicon and structure. The former becomes increasingly complex, but the structure/grammar needn't show such a development.*

*It's sometimes really difficult for SEs to see the difference between C1 and C2. The stacking of criteria isn't clear to everyone when you want to apply the criteria.*

*A change of level from KET to FCE takes time to adjust, which can make me occasionally less confident and extremely careful before I make final decisions.*

### *Comments on applying the descriptors: Interpreting standardisation samples*

While acknowledging the usefulness of PSN, some SEs suggested that they would appreciate having access to a wider range of performances and more guidance in the assessment commentaries.

*The problem with CAE is that there are insufficient examples of performance on PSN for SEs to be confident of what constitutes the highest grades.*

*I find that there are sometimes inconsistencies in the marks awarded when I watch the videos on the PSN – especially for Pronunciation and Discourse Management. This can result in some insecurity when assessing candidates in the exams.*

*I sometimes disagree with the marks awarded by Cambridge, especially for Pronunciation. If one rarely or never tests Asians, it's hard to assess the performance according to the criteria.*

*I think it's generally a very good system – they are clear. Watching PSN and reading the commentaries does a lot of good. Also not forgetting to look at the bullet points for each. Going on PSN and being reminded. Sometimes the commentaries on PSN could be a bit more elaborate. The system does work if you exploit it fully.*

### *Teaching experience and examining Cambridge English Speaking tests*

Proof of substantial, relevant, and recent teaching experience is one of the minimum professional requirements of prospective Speaking Examiners (Cambridge English Language Assessment 2013), so we can suppose that teaching is considered essential in being able to assess speaking performance. Question 1 of the survey asked Examiners to indicate which CEFR levels they teach (see Figure 1). There was no direct question linking teaching experience at a certain level with confidence in assessing at that level; however, five SEs spontaneously commented that their own confidence was greater if they had taught the level.

*I think I feel more confident examining the tests that I teach. Then I know the level, I know the tests, I know what's required.*

*I think it helps you be more confident if you teach the exam you're examining.*

*Because I teach more at the B2 level I find I have to refer even more to the criteria when judging BEC Higher.*

*If I lack confidence it's generally because I rarely teach and/or examine the level.*

*Since I teach university students and staff, I need to get into the levels, especially the lower ones. Recently I took on some beginners and this helps.*

### *The perceived helpfulness of various factors in building confidence*

The final section of the questionnaire focused on the training, support and operational stages of being an examiner, with the potential goal of identifying possibilities for improving SE training and development.

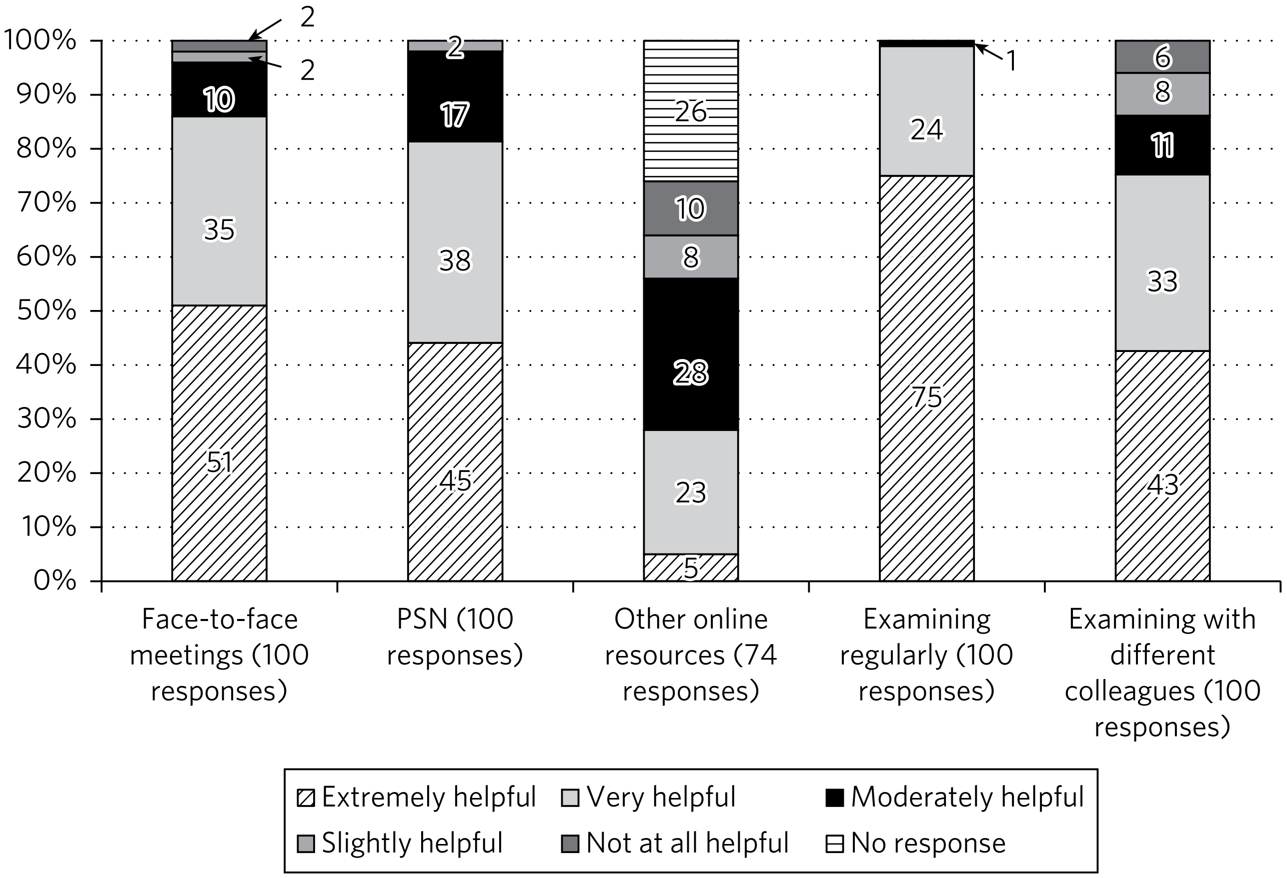
In addition to regular online (via the PSN) and face-to-face certification of procedure and assessment, SEs have permanent access via the PSN to online resources in the form of tests on video with assessment comments provided by Cambridge English. The other factors SEs could select as helpful in building confidence in rating were the use of other online materials (for example the Speaking test examples provided on the Cambridge English website for teachers), regular examining assignments and examining with a range of partner examiners. Responding SEs were asked to rate each of these factors on a scale of 'extremely helpful' to 'not at all helpful' (see Figure 7).

The feeling that examining regularly is 'extremely helpful' or 'very helpful' in gaining confidence in assessment is almost unanimous (99%). Face-to-face meetings and use of PSN online resources are perceived as almost equally useful, with respectively 86% and 83% of respondents identifying them as 'extremely helpful' or 'very helpful'. Fewer examiners, although still a clear majority (76%), see examining with different colleagues as 'extremely helpful' or 'very helpful'. A smaller number of SEs (28%) think other online resources are 'extremely helpful' or 'very helpful'. There were 74 responses to this question, compared with 100 for the others; three SEs commented that they did not know about any other online resources.

Although PSN and the face-to-face annual certification meetings are already seen as very useful tools in the ongoing QA and support of examiners, there were some thoughtful suggestions from SEs about how they could potentially be improved or refocused. The main focus of comments was on PSN:

*PSN: Watching the videos is very helpful especially just prior to examining but the comments which follow are often not helpful at all. They point out the positive points but give little help – i.e. useful informative comments – as to why a candidate received one mark rather than another. This can leave a less clear picture than*

**Figure 7: Responses to Survey Question 10: How helpful are the following in giving you confidence in making assessments?**



*before reading. [. . .] Perhaps it would be easier to go into more detail on the partial tests first – looking only at specific descriptors for one skill with a follow up whole test which exemplifies the critiques made beforehand?*

*PSN would be even more helpful if it contained more examples of exams and more students of different nationalities and backgrounds (i.e. European students vs Asian students).*

*PSN is helpful to refresh my memory after a long break. However, the candidates need to be changed regularly in order to avoid boredom and offer different conversations.*

*As a suggestion and perhaps to extend the use of PSN maybe three OEs could discuss their reasons for awarding different marks on a consensus basis and check with Cambridge.*

*It often seems that the candidates shown in Cambridge (videos) don't really correspond with the candidates assessed in Switzerland. Several of us find that we are assessing a larger percentage of C1 and C2 candidates who aren't really up to standard and perceive a difficulty to refine the marks awarded between 0 and 2.5.*

*Face-to-face meetings, and reflection on the relative value of different elements.*

*I'm aware of constraints [in certification meetings], but ideally, would like the sessions to be longer OR more frequent. Rather than exercises, I think more on-screen training with colleagues (assessing the grades for different levels) would be very useful. I feel that it is the combination of training, such as face-to-face and use of PSN, along with examining regularly that build confidence and enhance quality of outcome. At face-to-face meetings, however, there is often not enough time to thoroughly examine important issues and a lack of time for discussions. Also important is that all examiners have experience in both teaching and examining at the levels they examine, and that there are not discrepancies in understanding the descriptors and their application.*

## Discussion: Summary and conclusions

The fact that the response rate was high (70%) to a survey which was not obligatory and was sent in August (traditionally the holiday month in Switzerland) indicates that the issue of examiner confidence resonates among SEs in Switzerland. SEs indicated that some aspects of the complex process of rating speaking skills can cause them to feel at times less certain in awarding marks, and in addition they were able to identify possible areas where more support would be valuable to them in improving their expertise. This information could be a useful basis for improvements to professional development aimed at addressing perceived needs.

A clear majority of SEs feel 'confident' or 'very confident' in the Global and Analytical Assessments they make, and are satisfied with the training they receive, which is reassuring. However, factors relating to understanding and applying the criteria and concerned with managing the duality of the Interlocutor role were identified as possible causes for uncertainty in assessing, which suggests that there may be a need among some SEs for more focused training and professional development related to these areas:

- The rationale behind the Assessment Scales, and in particular why intermediate descriptors are not provided.
- Deepening understanding of the application of the Assessment Scales, in particular the Discourse Management Scale.
- Maintaining concentration throughout the test by:

– maintaining focus on assessment while dealing with the challenges of the Interlocutor role, including the time pressure of the test schedule and the difficulty of completing parts of some Speaking tests within the time allowed without making candidates feel rushed, as well as the requirement to judge when intervention may be necessary.

– maintaining focus on assessment when Interlocutor intervention in the test is necessary, in particular when the test is shorter and the language sample therefore smaller (e.g. *KET*).

- Making assessments where there is a more limited sample of language produced (*KET*, *BEC P*).

In addition, the following more general points came out of the study:

- Face-to-face certification meetings and PSN are both considered to be positive learning experiences by a large majority of SEs. However they made constructive and considered suggestions for improvement of both.
- Regular examining is felt by almost all SEs to be an important factor in gaining confidence in assessing performance, and a clear majority feel that examining with a range of colleagues is helpful.
- SEs were not directly asked about the value of their teaching experience. However, a small number of SEs spontaneously identified experience in teaching the level as a factor in increased understanding of the test level for assessments.

**Possible implications for SE development**

The question of becoming more confident in rating candidates appears to be one that SEs have reflected on individually and regard as desirable. It is, therefore, a reasonable professional response from examiner trainers to take this desire seriously and to develop ways of supporting examiners in this endeavour.

Comments made by some SEs suggest that during the training stage, more attention to deepening understanding of the scales and their interpretation across the CEFR levels would be valuable: expanding this aspect of training would increase the value of the subsequent reflective process, which is an aim of the annual certification process. Thought could be given to developing new approaches to this area of training and professional development.

Some SEs appear to need support in balancing the different aspects of the Interlocutor's role in some of the Speaking tests. Managing the test materials, script and procedure while arriving at a Global Mark require that the Interlocutor concentrate fully throughout each test. SE comments reflect how some of them experience this challenge as impacting on their assessments and this insight could be taken into account in future Speaking test revisions.

In their extended written comments or in subsequent interviews no SEs referred to the glossary or other information in the *Instructions to Speaking Examiners (ISE)* booklet, which could be of use to them in deepening their understanding of the criteria. Should there be some specific activities to encourage SEs to use the information in a more explicit way, both in the training process and in professional development materials?

Several SEs commented particularly on the challenge of understanding and assessing Discourse Management. Perhaps this criterion requires more explanation and practice in training sessions, in order to gain a fuller understanding. The other criteria – Grammar, Vocabulary, Pronunciation, Interactive Communication – can all be understood and used in the context of everyday teaching life. Discourse Management however is not a term in general use, so new SEs may not start with any understanding vof the concept, and in particular of how Discourse Management skills develop. If Discourse Management is the least accessible criterion for new examiners, then it must be ensured that all new examiners gain a clear understanding of it.

Some of the comments could suggest that SEs may be blurring the roles of teacher and examiner. Perhaps the differences in the roles should be an explicit point in examiner training so that the need to keep the roles separate can become clearer. Reminders in the ISE booklets or in professional development activities available for Team Leaders to use with SEs could be useful.

Developing and broadening the use of PSN was a point touched on by several SEs. Could the value of PSN be enhanced by expanding the range of activities available? SEs expressed their appreciation at being able to compare their performance with that of perceived 'expert examiners' during face-to-face meetings, and to discuss such issues as positioning, understanding and applying the scales. Activities replicating the training method of comparing your performance with that of a perceived expert could be devised and made available on PSN.

## Points for further investigation arising from this study

1. Where this group of SEs are 'occasionally unconfident' in making assessments, they sometimes relate this to particular Speaking tests. It might be useful to try to investigate further why some tests seem to be more challenging than others for SEs.
2. This was a small sample of SEs, meaning only very tentative conclusions can be drawn: it might be interesting to know whether SEs in other countries feel the same way as the Swiss SEs about the issue of confidence in assessing speaking skills.
3. Does confidence in one's ability to rate speaking performance impact on the ratings SEs make in live tests? For example:
   - Is there a relationship between degree of confidence and range of marks (0–5) awarded at different CEFR levels/

exams? Do more confident examiners use the scales differently from less confident ones?

- Does lack of confidence about applying any of the analytical criteria influence an SE's ability or willingness to use the full range of marks for that particular criterion?
- Is it possible to establish (perhaps by the use of statistical monitoring data already routinely gathered) whether SEs who feel more confident about rating are in fact more reliable markers?

In our experience of recruiting SEs, interpretation of what is *substantial and relevant* teaching experience has proved problematic, and it would be of interest to be able to gain more insight into this area. If teaching experience at a particular level could be shown to lead to greater confidence in one's assessments of candidates at that level, then it would indeed be an important requirement to consider when recruiting SEs for different Speaking tests.

## References

Anders Ericsson K, Prietula, M J and Cokely, E T (2007) The making of an expert, *Harvard Business Review*, 1–7.

Brown, A and Taylor, L (2006) A worldwide survey of examiners' views and experience of the revised IELTS Speaking test, *Research Notes* 26, 14–18.

Cambridge English Language Assessment (2013) *Guidelines for Speaking Examiner Certification*, Cambridge: Cambridge English Language Assessment.

Chambers, L, Galaczi, E and Gilbert, S (2012) Test taker familiarity and speaking test performance: Does it make a difference? *Research Notes* 49, 33–40.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Humphry-Baker, A (2000) *Speaking tests: Students' perception and performance*, unpublished MA thesis, University of Manchester.

Moon, J (2005) *The Reflective Teacher*, available online: www.mcgraw-hill.co.uk/openup/chapters/9780335222407.pdf

Ofsted (2004) *Why Colleges Succeed*, available online: www.ofsted.gov.uk/resources/why-colleges-succeed

Yates, L, Zielinski, B and Pryor, E (2008) The assessment of pronunciation and the new IELTS pronunciation scale, *IELTS Research Reports Volume 12*, available online: www.ielts.org/researchers/research/volume_12.aspx

## Appendix 1

### Survey questions

1. Which CEFR levels do you teach?
2. How long have you been a Cambridge English Speaking Examiner?
3. Which Main Suite and professional exams are you certified for?
4. On the whole, which examiner role do you feel more confident in?
5. As Interlocutor, how confident do you feel awarding Global Marks in each exam?  
Very confident/confident/occasionally unconfident/not confident/N/A
6. As Interlocutor, what factors can limit your confidence when awarding Global Marks? (Please consider each exam separately. More than one answer possible.)  
Having both to manage procedure and award marks/relating performance to descriptors/absence of expanded descriptors for some bands/lack of training/lack of experience/N/A
7. As Assessor, how confident do you feel when awarding Analytical Marks in each exam?  
Very confident/confident/occasionally unconfident/not confident/N/A
8. As Assessor, what factors can limit your confidence when awarding Analytical Marks? (Please consider each exam separately. More than one answer possible.)  
Having both to manage procedure and award marks/relating performance to descriptors/absence of expanded descriptors for some bands/lack of training/lack of experience/N/A
9. As Assessor, how confident do you feel applying each of the analytical criteria?  
Very confident/confident/occasionally unconfident/not confident/N/A
10. How helpful are the following in giving you confidence in making assessments? (Face-to-face meetings, PSN, Other online resources, Examining regularly, Examining with different colleagues)  
Extremely helpful/very helpful/moderately helpful/slightly helpful/not at all helpful
11. Where do you examine? Please indicate Centre number(s).
12. Age
13. Gender
14. What is your native language?
15. Examiner number (optional – for research purposes only)

# Cambridge English Funded Research Programme Round 6: Call for Proposals

**Cambridge English is once again making grant funding available for research projects to be conducted during 2015.**

Educational institutions and qualified researchers are invited to submit research proposals on topics related to the following Cambridge English examinations and teaching qualifications:

**General English and for Schools**

| | | | | |
|---|---|---|---|---|
| *Cambridge English: Starters (YLE Starters)* | *Cambridge English: Movers (YLE Movers)* | *Cambridge English: Flyers (YLE Flyers)* | *Cambridge English: Key (KET)* | *Cambridge English: Key (KET) for Schools* |
| *Cambridge English: Preliminary (PET)* | *Cambridge English: Preliminary (PET) for Schools* | *Cambridge English: First (FCE)* | *Cambridge English: First (FCE) for Schools* | |

**Academic and Professional English**

| | | |
|---|---|---|
| *Cambridge English: First (FCE)* | *Cambridge English: Advanced (CAE)* | *Cambridge English: Proficiency (CPE)* |
| *Cambridge English: Business Preliminary (BEC Preliminary)* | *Cambridge English: Business Vantage (BEC Vantage)* | *Cambridge English: Business Higher (BEC Higher)* |
| *Cambridge ILEC (International Legal English Certificate)* | *Cambridge ICFE (International Certificate in Financial English)* | |

**Teaching Qualifications**

| | | | |
|---|---|---|---|
| *TKT (Teaching Knowledge Test)* Core modules:<br>• TKT: Module 1<br>• TKT: Module 2<br>• TKT: Module 3<br>• *TKT: Practical* | *TKT (Teaching Knowledge Test)* Specialist modules:<br>• *TKT: CLIL*<br>• *TKT: Knowledge About Language*<br>• *TKT: Young Learners* | CELTA (Certificate in Teaching English to Speakers of Other Languages) | Delta Modules (Diploma in English Language Teaching):<br>• Delta Module One<br>• Delta Module Two<br>• Delta Module Three |

**Research topics of interest**

All research proposals should relate to one of the following research areas (1-4) and sub-topics (a-o):

Areas:

1. Test validation
2. Contexts of test use
3. Test impact
4. Learning-oriented assessment

Sub-topics:

a. Cognitive processes of test takers
b. Writing, Speaking, Listening or Reading test features that distinguish exams or test taker performance at adjacent proficiency levels
c. The impact of different first language on test performance at adjacent proficiency levels
d. The nature of General/Academic/Professional English language ability
e. Different types of validity of a test for a specific group of candidates
f. The impact of different task types on candidate performance
g. The impact of delivery mode on test performance
h. The impact of test accommodations on test taker performance
i. Comparability studies that relate a Cambridge English test to another in a specific context
j. Perceptions towards General/Academic/Professional Cambridge English exams among stakeholders, e.g. those involved in test preparation
k. Perceptions towards Cambridge English Teaching Qualifications among stakeholders.
l. English language progression for non-native speakers in tertiary contexts
m. Test impact on English language programmes (e.g. hours of study and score gains)
n. The impact of assessment on language learning, teaching practices and institutional change.
o. The nature of content knowledge of teaching in contexts where English is used as a medium of instruction

Consideration will also be given to other issues of current interest in relation to General, Academic or Professional Cambridge English examinations and our teaching qualifications.

**Level and duration of funding**

Successful proposals will receive funding of up to £2,000, £5,000 or £15,000 including institutional overheads. Project duration is for ten months and the research undertaken will be covered by a research agreement, to be signed before research can commence.

**Schedule for Round 6**

| | |
|---|---|
| **01 August 2014** | Call for proposals |
| **01 October 2014** | Deadline for submission of proposals |
| **17 October 2014** | All applicants notified |
| **November 2014** | Successful applicants sign research agreement and finalise design |
| **01 December 2014** | Research projects commence |
| **01 May 2015** | Interim reports due |
| **01 October 2015** | Final reports due |

**Submission of proposals**

Please download and complete the Proposal Form fully and note that late or incomplete submissions will not be accepted. You can download the form from www.cambridgeenglish.org/research-and-validation/research-and-collaboration

All applications and queries should be sent to validation@cambridgeenglish.org with the subject 'CEFRP Round 6'.

**Evaluation criteria**

The Funded Research Committee will evaluate all complete proposals using the following criteria:

1. Relevance of topic to Cambridge English and contribution to knowledge more widely
2. Sound theoretical basis and evidence of a rigorous research design
3. Relevance of researchers' experience
4. Potential for practical outcomes
5. Potential for publication

All applicants will be notified of the outcome of their submission.

Research

To subscribe to *Research Notes* and download previous issues, please visit:
**www.cambridgeenglish.org/research-notes**

## Contents:

For further information visit the website:
**www.cambridgeenglish.org**

Cambridge English
Language Assessment
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

**www.cambridgeenglish.org/help**

ALTE Association of Language Testers in Europe

A DIVISION OF
CAMBRIDGE ASSESSMENT