ResearchNotes

Contents

The URL for reading/downloading single articles or issues of *Research Notes* is:
www.CambridgeESOL.org/rs_notes

The URL for subscribing to *Research Notes* is:
www.CambridgeESOL.org/rs_notes/inform.cfm

Editorial Notes

Welcome to issue 20 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The theme of this issue is impact, that is the effects that our assessment products have on a diverse range of stakeholders worldwide, including candidates, examiners, teachers, institutions and governments. Impact can take many forms, from the introduction of a new exam or changing a test format, via revising existing mark schemes, to the more localised 'washback' into the classroom setting or to an organisation's decision to use a test for a specific purpose. Supporting the wider research community by funding research projects and through publications can also be considered impacts.

In the opening article Lynda Taylor discusses notions of washback and impact, locating them within the broader framework of consequential validity. She highlights the long tradition of consultation which we have enjoyed with our test stakeholders and the more recent role of systematic impact studies within our approach to test development and validation.

The next two articles describe trialling new or revised language tests which is a major part of assessing their likely impact on candidates and examiners. Louise Maycock and Tony Green report the results of a study which investigated whether computer familiarity affected candidates' computer-based IELTS test scores and also addressed candidates' attitudes towards this new format which is available from May 2005. Next, Stuart Shaw and Evelina Galaczi describe a trial of mark schemes for the forthcoming Skills for Life tests which aimed to improve the assessment process and refine the rating scales for the writing component. This project was undertaken to ensure that assessment was standardised, ensuring a fair result for all candidates taking these modular tests designed to improve adult literacy in the UK.

The following articles are concerned with investigating candidates' spoken production in general and academic English contexts. Firstly, John Read (University of Wellington, New Zealand) reports on an IELTS funded study which explored vocabulary use by candidates in the revised IELTS speaking test across a range of band scores; the findings provide validation evidence for the revised speaking assessment criteria and scales introduced in 2001. Next, Evelina Galaczi presents performance data for FCE, CAE and CPE Speaking Tests for 2003, reporting on work to validate these Upper Main Suite tests through analysis of scoring criteria and examiner behaviour.

Roger Hawkey then reports on a study which explored the washback effects of the revised Certificate of Proficiency in English (CPE) on textbooks. Next we describe Cambridge ESOL's contribution to a report on the teaching of languages to learners with special needs and we provide details of several recent publications of potential interest to the language testing community.

We end this issue with calls for proposals for the 11th Round of the IELTS Funded Research Program and for Masters level dissertations on language testing which can be submitted for the IELTS Masters Award 2005.

Washback and impact: the view from Cambridge ESOL

LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

Washback and impact: some definitions

Alderson and Wall (1993) suggested that the term 'washback' provides a useful metaphor to help us explore the role of language tests in teaching and learning, i.e. in relation to factors such as the individual learner, the teacher's attitudes and behaviour, the classroom environment, the choice and use of teaching/learning materials. 'Washback' allows for the possibility of effects of tests on teaching to be viewed on a continuum – stretching from negative (harmful) at one end, through neutral, and into positive (beneficial) at the other end. Negative washback is said to occur when test content or format is based on a narrow definition of language ability and so constrains the teaching/learning context. Positive washback is said to result when a testing procedure encourages 'good' teaching practice and positive learning outcomes.

The past ten years have seen a growing awareness that language testing also has consequences beyond the immediate teaching/learning context. The use of tests and test scores can impact significantly on the career or life chances of individual test-takers (e.g. access to education/employment opportunities). They can also impact on educational systems and on society more widely: for example, test results are used to make decisions about school curriculum planning, or funding allocation; about immigration policy, or licensing for health professionals such as doctors and nurses. Widespread acceptance of a test often encourages publishers to produce test preparation materials and motivates institutions to run preparation courses. These wider consequences are often referred to as 'test impact' (Bachman and Palmer 1996).

Some language testers consider washback as one dimension of impact, describing effects on the educational context; others see washback and impact as separate concepts relating respectively to 'micro' and 'macro' effects within society. Most language testers locate both concepts within the theoretical notion of 'consequential validity' in which the social consequences of testing are part of a broader, unified concept of test validity (Messick 1996). Consequential validity – along with related themes of fairness and ethics – has been extensively discussed in recent years (see Kunnan 2000) and most language testers now acknowledge washback and impact to be highly complex phenomena requiring systematic investigation.

The stakeholder dimension

Underlying these complex phenomena is an equally complex network of relationships among the many different 'stakeholders' who populate the world of language teaching and testing. It is obvious that learners and teachers are directly affected by the washback of a language test, but other stakeholders on whom a test can impact include parents, employers, teacher-trainers, researchers, school owners, test writers, publishers, and examiners. Different tests have different stakeholder constituencies with multiple voices all needing to be heard; direct consultation with members of the stakeholder community is one way of gauging the nature and extent of a test's influence.

Historically, there has always been a close relationship between Cambridge ESOL's language tests and the world of English language teaching; this is described in detail in recently published accounts of the development of the CPE and CELS examinations (Weir and Milanovic 2003, Hawkey 2004). As a result, our test development and revision methodology has always had built-in opportunities for direct consultation with key stakeholders such as teachers and materials writers who also serve as item writers and examiners; in addition, direct contact with our network of test centres worldwide and with those organisations who use our test scores for decision-making purposes means we can gather ongoing feedback on test usefulness. Formal stakeholder surveys during review/revision projects for established tests allow us to identify specific issues which may need addressing, and product/market surveys help shape new test developments to meet evolving needs. Our teacher seminar programme and teaching resources websites have provided new communication channels which enhance the consultation process and help monitor test impact.

Researching aspects of washback and impact: the Cambridge ESOL approach

Cambridge ESOL's tradition of engaging with the stakeholder community reflects a long-standing concern to achieve washback/impact that is as positive as possible; over the past 10 years, however, we have sought to develop a more formal and systematic approach to researching the nature of washback/impact as an integral part of the test development and validation process. It was for this reason that Cambridge ESOL initiated a project in the mid 1990s to develop suitable instruments for investigating specific research questions relating to the Cambridge tests (Milanovic and Saville 1996). In collaboration with researchers at Lancaster University, a set of research tools were developed and validated between 1996 and 1999; these have since been used in a number of specially designed impact studies to investigate: the attitudes and perceptions of test-takers; the attitudes and perception of other key stakeholders, e.g. teachers and administrators; the nature of test preparation courses and materials; the nature of classroom activity. Formal impact studies are especially important where high-stakes tests are concerned so it is not surprising that much attention has focused on IELTS to complement other impact-related studies conducted under the joint-funded IELTS research program; however, Cambridge ESOL has also carried out studies relating to the impact of Main Suite exams (e.g. the CPE Textbook Washback study; the PL2000 Impact study).

Cambridge ESOL frequently undertakes or supports other activities which could be regarded as impact-related, especially in

relation to high-stakes tests such as IELTS. In 1999, for example, rapid growth in the use of IELTS General Training for immigration purposes prompted Cambridge ESOL to conduct a review and consultation process with language testing experts in the UK, Australia and New Zealand to ensure the test's suitability for this purpose. More recently, the increased use of IELTS by professional licensing bodies led in 2004 to two major standard-setting studies being conducted with professional bodies in the UK and US who are responsible for licensing health service personnel; a similar standard-setting study is planned in 2005 with the Canadian immigration services. The purpose of such studies is to ensure that a test is 'fit for purpose' and that the choice of decision-making cut-scores is sound and defensible.

A concern for the social consequences of test use is a mark of ethical language testing (Hamp-Lyons 1997); it reflects an acknowledgement that individuals and society have expectations of good value and accountability, as well as a commitment to act in a manner that is professionally and socially responsible. As early as 1994 Cambridge ESOL and its ALTE partners drew up and adopted a Code of Practice – acknowledging certain obligations and making explicit the standards of quality and fairness adhered to.

Today tests are increasingly used for 'high-stakes' gate-keeping and policy-making purposes, as well as to provide targets for and indicators of change; it is therefore even more important for test producers to provide appropriate evidence that the social consequences of using their tests in such ways is beneficial rather than detrimental. Until fairly recently, claims and assertions about the nature and extent of a test's impact were largely based upon impression and assumption. It was relatively simple for producers to claim positive washback for their own tests, or for users to criticise tests on the grounds of negative washback; but it was also too easy for both sets of assertions to go unchallenged. Impact research – such as that conducted by our own organisation – reflects the growing importance of evidence-based approaches to education and assessment which enable policy and practice to be justified in terms of sound evidence about their likely effects.

Cambridge ESOL's contribution to this field is not insignificant. Over the past 10 years we have been able to develop and refine a suitable methodology for the empirical investigation of washback/impact; findings from our washback/impact studies have been widely reported at conferences and published in the literature (Saville and Hawkey 2004, Hawkey forthcoming); and the systematic investigation of washback/impact has been integrated into our model for test development and validation.

References and further reading

Alderson, J and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 116–29.

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

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

Hawkey, R (2004) *A Modular Approach to Testing English Language Skills*, Studies in Language Testing, Vol 16, Cambridge: UCLES/Cambridge University Press.

—(forthcoming) *The theory and practice of impact studies: Messages from studies of the IELTS test and Progetto Lingue 2000*, Studies in Language Testing, Vol 24, Cambridge: UCLES/Cambridge University Press.

Kunnan, A (2000) *Fairness and validation in language assessment: selected papers from the 19th Language Testing Research Colloquium*, Studies in Language Testing, Vol 9, Cambridge: UCLES/Cambridge University Press.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13/4, 241–256.

Milanovic, M and Saville, N (1996) *Considering the impact of Cambridge EFL Examinations*, Internal working report, Cambridge: University of Cambridge Local Examinations Syndicate.

Saville, N and Hawkey, R (2004) The IELTS Impact Study: Investigating Washback on Teaching Materials, in Liying Cheng, and Watanabe, Y (Eds) *Washback in Language Testing: Research Contexts and Methods*, New Jersey: Lawrence Erlbaum Associates.

Taylor, L (2005) Key concepts in ELT: washback and impact, *ELT Journal*, 59/2, 154–155.

Weir, C and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing, Vol 15, Cambridge: UCLES/Cambridge University Press.

The effects on performance of computer familiarity and attitudes towards CB IELTS

LOUISE MAYCOCK AND **TONY GREEN**, RESEARCH AND VALIDATION GROUP

Introduction

CB IELTS is a linear computer-based (CB) version of IELTS that has been under development since 1998 and will shortly be made available to candidates as an alternative to the paper-based (PB) test, initially at a limited number of centres. Prior to its release, CB IELTS has been the subject of extensive user testing and trialling, culminating in a trial conducted world-wide under live test conditions, the first phase of which was reported in *Research Notes* 18, focusing on the comparability of test scores obtained through the two formats.

In addition to the question of comparability of results, levels of computer familiarity and their potential impact on test scores is also of concern. Eignor et al (1998), Kirsch et al (1998) and Taylor et al (1998) explored the effects of computer familiarity on

computer-based TOEFL scores. Taylor et al (1998) found that when candidates were required to view an introductory tutorial before sitting the test, there were no meaningful differences between the test scores obtained by computer familiar and non-familiar candidates. CB IELTS is also administered in conjunction with an introductory tutorial and sample materials are provided so that candidates can become familiar with the CB test. O'Sullivan et al (2004) in an investigation of writing on computer-based and paper-based versions of IELTS also found no link between computer familiarity and test performance. However, they caution that, 'although the evidence points to the view that computer familiarity alone may not have a significant effect on CB test results… it cannot be ignored when comparing [PB] and [CB] tests' (O'Sullivan et al 2004:9).

All candidates who took part in the CB IELTS trial also completed a questionnaire addressing computer familiarity and attitudes towards the CB test (see Appendix 1 on page 7). This paper reports on the responses to this questionnaire in relation to two key research questions:

- What are the attitudes of candidates towards CB IELTS?
- How do candidates' computer experience and ability, and their attitudes towards features of CB IELTS as measured by the questionnaire influence their performance on the CB test in relation to their performance on the PB test?

Participants and instruments

Participants included 882 candidates, just over half of whom (50.8%) spoke Chinese as a first language. 56.6% were women and most (72%) were aged between 19 and 29 (see Table 1).

Table 1: Candidates participating in CB IELTS trial

Gender	**Per cent**
Female	56.6
Male	43.4
Age	**Per cent**
16-18	8.8
19-22	32.1
23-29	39.9
30-39	15.0
40 and above	4.2

The 26-item questionnaire was developed by the Research and Validation Group and addressed the following areas (the questions with percentage responses are provided in Appendix 1):

- Using writing tools in the CB environment: Questions 11,12,13
- Timing on the objective CB components: Q2,4,8
- Timing on the objective PB components: Q3,5,9
- Reading text on screen: Q7
- Timing on the CB Writing test: Q14
- Timing on the PB Writing test: Q15
- Experience and ability with computers: Q18,19,20,21
- Perceived value of the Specimen Materials CD-ROM (provided in advance of the test administration): Q16
- Perceived value of the introductory CB tutorials (provided at the beginning of the test administration): Q1,6,10
- Preference for the CB or PB test: Q22,24
- Perceived value of computer ability for CB success: Q17
- Previous experience of IELTS: Q23

In an open-ended question, respondents were also invited to make comments on the CB test: 'If you wish to comment further on any of the above statements, or any other aspect of CB IELTS, please do so in the box below.' These comments were collated and analysed for key themes.

Findings

Using writing tools on the computer

The CB Writing component offers candidates the option of composing their responses to the two tasks on screen using the keyboard or on paper, writing by hand. Items Q-11, Q-12 and Q-13 addressed the editing and word count functions provided to candidates who opted to respond to the writing test on screen.

These text editing features were generally seen as both a user friendly and a helpful feature of the CB test. The 24% of candidates who responded, 'Don't Know/Not Applicable' to Q-11, Q-12 and Q-13 reflects the percentage of candidates who chose to respond on paper and so did not need to use these features. Of those who responded with the other options, 77% agreed that the editing tools were easy to use (Q-11) and 68% agreed that they were useful (Q-12). This compares with 17% (Q-11) and 24% (Q-12) who disagreed with the statements.

Timing

For Listening, candidates were generally happier with the timing on the PB test. 58.2% disagreed with the statement 'On the computer-based test, I had enough time to review my answers at the end of each section' (Q-4), 30.8% selected 'disagree', 27.3% 'strongly disagree'. This compares with figures of 11.6% and 4.4% respectively in response to the parallel statement referring to the PB Listening component (Q-5).

These results were associated with a difference between the CB and PB versions of the test. In the PB test, candidates need to transfer their answers from the question paper to a computer-readable optical mark sheet to allow for machine scoring of responses. This is unnecessary for the computer-based test and no additional time was provided. Comments from candidates made it clear that some felt disadvantaged by not being allowed this time for answer review, as they misconceived it, on the CB test as well. Further investigations were carried out to establish whether candidates were disadvantaged by the absence of the ten minute transfer period and these established that there was no significant effect on scores. The CB Listening test will therefore not include a ten minute transfer period.

For Reading, timing was generally less of an issue on the CB test (Q-8, Q-9), but here the differences were less clear cut than on the Listening component. On the Writing component there seemed to

be little difference between the formats. Twenty per cent of those selecting responses other than 'Don't know/Not applicable' to Q-14 (relating to the CB test) and 23% of those responding to item Q-15 (the PB test) disagreed that they had enough time to complete both writing tasks in an hour.

Reading on screen

As set out in the questionnaire responses in Appendix 1, candidates were almost evenly split on whether they agreed or disagreed that long texts were more difficult to read on screen than on paper (15% strongly agreed that the texts were more difficult to read on computer and 13% strongly disagreed). This suggests, encouragingly, that the CB texts are no more difficult to read than the PB texts for most candidates and that overall there is no clear advantage in one mode of presentation over the other. The introduction of CB IELTS should provide candidates with a welcome opportunity to choose the format that they prefer.

Support materials

The reaction to the introductory tutorials and the specimen materials was generally very positive. The proportion agreeing that the materials had been useful was relatively lower for the introductory tutorial to the Writing module (Q-10) and the general specimen materials CD-ROM (Q-16) than for the introductory tutorial for the Listening and Reading components (Q-1 and Q-6). However, in both cases there were large numbers of 'Not applicable' responses, which probably represent those candidates who completed the Writing CB paper by hand. Only a very small proportion of the candidates (between 2% and 5%) disagreed that the introductory tutorial gave sufficient information on how to answer questions in any test section (Q-1, Q-6, Q-10) and just 3% disagreed that the specimen materials CD-ROM had helped them know how to do the test (Q-16).

Computer ability and experience

Most candidates were confident computer users with some experience of composing written responses on the computer. They generally believed that knowledge of computers would be an advantage when taking CB IELTS: 67.7% agreed that candidates with good computer skills would perform better on CB IELTS than those with only basic skills (Q-17). Among candidate comments collected through the open-ended 'comments' box, it was generally perceived, even by those who preferred the CB test, that doing the IELTS test on computer advantaged those with good computer or keyboard skills.

The majority of candidates felt able to use a computer. The greatest proportion rated their computer ability to be 'intermediate' (44%). However, a range of skill levels were represented with 31% rating themselves as 'basic' and 20% as 'advanced' users. Very few candidates (1%) rated their ability as 'zero'. Thirty three per cent of candidates who responded to Q-21 and did not choose 'Don't know' felt that they were faster typists than writers with 48% believing they were faster at handwriting and 19% claiming to be the same speed. Use of computers was most frequent 'at home' (Q-19) with 64.9% claiming to use computers 'often', followed by use 'at work/ school' (45.8% 'often'). Most also claimed to use computers at least 'sometimes' for writing essays (78.5%) with 15.4% selecting 'never'.

Other issues

Almost half of the sample had taken IELTS before and so would already have been familiar with the PB format. Just over half of candidates preferred CB IELTS to the paper-based version and would choose it over the PB test in the future. The largest proportion (41% of those who responded) had preferred taking the CB test to the PB version with 35% preferring the PB test and 24% expressing no preference.

Relationships between questionnaire scales

Over half of the correlations between questionnaire scales were significant ($p<.01$), but these were generally between 0.1 and 0.4 indicating modest relationships (see Table 2). The strongest correlations were between experience and ability with computers and management of the CB IELTS editing tools and word count features. There was also a relatively strong relationship between computer experience and having enough time to complete the Writing test on the computer, which was in turn also related to using the text editing and word count features of CB IELTS. In short, as might be expected, candidates with more experience of computers felt better able to exploit the word-processing tools on offer.

Interestingly, those who agreed that the introductory tutorial and, to a lesser extent, the specimen materials CD-ROM were helpful were also likely to find value in the editing tools, suggesting that some candidates had gained awareness of these tools from the support materials. A preference for the CB test over the PB version was also moderately related to experience with computers, management of writing tools, satisfaction with the timing of the CB test and a preference for reading on screen.

Analysis of Covariance

A shortcoming of the correlational analysis described above is that questionnaire factors are correlated to some extent with each other. They also display variation between groups based on first language or age. This will affect the interpretation of the results because the impact of one factor or background characteristic may mask the impact of another on test scores.

Analysis of Covariance (ANCOVA) enables us to factor out the influence of other variables when investigating the relationship between test performance and features of interest. In this case three separate analyses were carried out, on each of the Reading, Writing and Listening components. In addition to the questionnaire items, the variables used included the following:

- **Dependent variables:** performances on each of three components: Reading, Writing and Listening (Speaking, the fourth component of the test, is identical for CB and PB versions and so was excluded as a dependent variable from this study).
- **Independent variables:** First language (Chinese and non-Chinese); Gender; Age range (ages 15 and under;

Table 2: Correlations between questionnaire scales

	Exp. and ability with computers	Using writing tools in CB	Timing on CB objective components	Timing on PB objective components	Reading text on screen	Timing on CB Writing	Timing on PB Writing	Specimen materials CD-ROM	Introductory tutorial	Preference for CB or PB	Value of computer ability
Using writing tools in CB	0.42										
Timing on CB objective components	0.34	0.32									
Timing on PB objective components	0.12	0.03	0.25								
Reading text on screen	-0.13	-0.16	-0.19	0.08							
Timing on CB Writing	0.44	0.45	0.40	0.15	-0.09						
Timing on PB Writing	0.01	0.06	0.17	0.34	0.02	0.34					
Specimen materials CD-ROM	0.18	0.31	0.28	0.10	-0.08	0.20	0.09				
Introductory tutorial	0.23	0.46	0.34	0.12	-0.13	0.23	0.09	0.46			
Preference for CB or PB	0.30	0.24	0.29	-0.05	-0.21	0.22	-0.03	0.13	0.12		
Value of computer ability	-0.11	-0.05	-0.05	0.01	0.18	-0.07	0.01	0.11	0.04	-0.05	
Previous experience of IELTS	-0.02	-0.02	-0.01	0.07	0.08	0.03	0.07	0.00	0.03	-0.08	0.10

16–18; 19–22; 23–29; 30–39; 40 and above); Previous experience of IELTS.

- **Covariates:** scores on the three PB components and on Speaking; Using writing tools in the CB environment; Timing on the objective CB components; Timing on the objective PB components; Reading text on screen; Timing on the CB Writing test; Timing on the PB Writing test; Experience and ability with computers; Perceived value of the specimen materials CD-ROM; Perceived value of computer ability for CB success; Perceived value of the introductory CB tutorials; Preference for the CB or PB test.

The results of the ANCOVA revealed just two significant ($p<.01$) main effects for variables other than PB scores: Age and Previous experience of IELTS were both found to have an impact on Listening scores. Additionally, there was a significant interaction between Age, Gender, First language and Previous experience of IELTS on the Listening component. However, the impact of these features on scores (as revealed by the eta squared statistic) was minimal. This suggests that, for these self-selecting candidates, scores on the CB test are little affected by differences of background or the issues addressed in the questionnaire regarding the CB format.

Conclusion

CB IELTS is generally popular with candidates and the preparatory materials (i.e. introductory tutorials and specimen materials) were found to be useful. Candidates who chose to compose their responses using the keyboard generally found no problems in managing the writing tasks on computer and found the functions available to them (cut, copy, paste and word count) useful.

The candidates taking part in the trial were reasonably confident in their own ability to use computers, but the majority felt that candidates with more advanced computer skills would perform better on CB IELTS than those with only basic skills. However, this was not borne out in the analysis. In common with results from similar studies (O'Sullivan et al 2004, Taylor et al 1998), candidate ability and experience in using computers was not found to have any significant impact on the difference between PB and CB scores for any of the tested skills.

Candidates expressed some dissatisfaction with the timing of the Listening component of the CB test in relation to the PB test. However, this dissatisfaction was not reflected in any significant difference in scores. Comments from candidates suggest that the dissatisfaction could be attributed to the absence in the CB Listening of the ten minute transfer time provided at the end of

Appendix 1: CB IELTS questionnaire responses

	Strongly agree/ Agree %	Neither agree nor disagree%	Disagree/ Strongly disagree %	Missing %
THE LISTENING MODULE				
Q-1: On the COMPUTER-BASED test, the introductory tutorial gave me enough information on how to answer the Questions.	90.8	3.9	4.5	0.8
Q-2: On the COMPUTER-BASED test, I had enough time to read the Q-s at the start of each section.	54.5	17.6	26.9	1.0
Q-3: On the PAPER-BASED test, I had enough time to read the Questions at the start of each section.	65.2	16.2	17.3	1.3
Q-4: On the COMPUTER-BASED test, I had enough time to review my answers at the end of each section.	24.7	13.2	58.2	3.9
Q-5: On the PAPER-BASED test, I had enough time to review my answers at the end of each section.	67.8	14.2	16.0	2.0
THE READING MODULE				
Q-6: On the COMPUTER-BASED test, the introductory tutorial gave me enough information on how to do the test.	93.0	3.9	2.6	0.5
Q-7: I found the long texts more difficult to read on screen than on paper.	39.8	20.1	38.2	1.9
Q-8: On the COMPUTER-BASED test, I had enough time to finish the whole paper in one hour.	63.5	16.0	19.2	1.3
Q-9: On the PAPER-BASED test, I had enough time to finish the whole paper in one hour.	43.5	22.2	32.8	1.5
THE WRITING MODULE				
Q-10: On the COMPUTER-BASED test, the introductory tutorial gave me enough information on how to do the test.	75.7	6.7	1.8	15.8
Q-11: On the COMPUTER-BASED test, I found the cut, copy, paste and undo functions easy to use.	58.4	13.3	4.4	23.9
Q-12: On the COMPUTER-BASED test, I found the cut, copy, paste and undo functions useful.	51.7	18.3	6.2	23.8
Q-13: On the COMPUTER-BASED test, I found the word count function useful.	63.4	9.8	2.7	24.1
Q-14: On the COMPUTER-BASED test, I had enough time to complete both tasks in one hour.	44.6	17.2	15.8	22.4
Q-15: On the PAPER-BASED test, I had enough time to complete both tasks in one hour.	51.7	19.6	21.3	7.4
IN GENERAL				
Q-16: The specimen materials CD-ROM helped me to know how to do the COMPUTER-BASED test.	84.2	9.9	3.1	2.8
Q-17: I think people with good computer skills will do better at CBIELTS than those with basic computer skills.	67.7	16.3	13.8	2.2
Q-18: In general, I am confident of my ability to use a computer.	63.3	18.8	16.0	1.9
	Never	**Sometimes**	**Often**	**Missing**
Q-19: How often do you use a computer:				
a) at home?	3.2	26.6	64.9	5.3
b) at work/school?	4.6	43.8	45.8	5.8
c) to write essays?	15.4	44.9	33.6	6.1
	Zero	**Basic**	**Intermediate**	**Advanced**	**Missing**
Q-20: How would you rate your level of computer ability?	1.2	31.1	43.7	19.6	4.4
		No	**The same**	**Yes**	**Missing**
Q-21: I can type faster than I can write by hand.		40.4	15.9	28.2	15.5
Q-22: I preferred taking the COMPUTER-BASED test to the PAPER-BASED test.		33.1	23.1	38.7	5.1
		No	**Yes**	**Missing**	
Q-23: I have taken IELTS before.		50.0	45.0	5.0	
		CB	**PB**	**Missing**	
Q-24: Given a choice in the future, which would you choose?		53.3	41.7	5.0	

the PB Listening component. However, investigation has shown that candidates do not receive any benefit in terms of score increases from receiving the transfer time on the PB component.

There is evidence that candidate background (age range, first language, gender and previous experience of IELTS) had some influence on test scores, but the effects were minimal and would have no meaningful impact on results. However, this is an area which will be made the focus of further research once more data from CB IELTS candidates becomes available.

The opportunity for candidates to choose between composing their responses on paper or on screen is seen as an essential feature of CB IELTS and allows them to select the response mode that reflects their usual practice. In this way, all candidates should have the opportunity to perform to the best of their ability, whether they are more accustomed to composing on computer or on paper. In the next issue of *Research Notes* Andy Blackhurst will report the results of CB IELTS Trial B together with the continuing validation work involved in the roll-out of computer-based IELTS to a number of venues from May 2005 onwards.

References and further reading

Eignor, D, Taylor, C, Kirsch, I and Jamieson, J (1998) Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Examinees, TOEFL Research Report 60, Princeton, NJ: Educational Testing Service.

Kirsch, I, Jamieson, J, Taylor, C, and Eignor, D (1998) Computer familiarity among TOEFL examinees, TOEFL Research Report 59, Princeton, NJ: Educational Testing Service.

O'Sullivan, B, Weir, C and Yan, J (2004) Does the computer make a difference?, unpublished IELTS Research Project Report.

Russell, M (1999) Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper, Education Policy Analysis Archives, 7:20.

Taylor, C, Jamieson, J, Eignor, D and Kirsch, I (1998) The relationship between computer familiarity and performance on computer-based TOEFL test tasks, TOEFL Research Report 61, Princeton, NJ: Educational Testing Service.

Skills for Life writing mark scheme trial: validating the rating scale for Entry Levels 1, 2 and 3

STUART D SHAW AND **EVELINA GALACZI**, RESEARCH AND VALIDATION GROUP

Introduction

Skills for Life is the UK's national strategy for improving adult literacy, numeracy and ESOL skills. In support of this strategy, Cambridge ESOL is working on new modular tests in ESOL Skills for Life at Entry Levels 1, 2 and 3, Level 1 and Level 2, for use by colleges and other learning providers. These will be at the same level as the other Cambridge ESOL exams, but have been specifically developed to meet the needs of UK residents who would like to develop their language skills.

Adoption of a new test requires that appropriate validation studies are conducted before the test is implemented in the live operational context. Significant features of a validation programme, based on the Cambridge ESOL model of test development (most recently described in Saville 2003), are the prominence of the criteria of the communicative construct, and the espousal of rigorous quantitative and qualitative systems for the establishment of validity, reliability, impact and practicality.

This article reports on a validation trial conducted on the Skills for Life (SfL) writing mark schemes at Entry Levels 1, 2 and 3. The trial entailed multiple re-rating of scripts by a team of independent examiners. In outline, the procedure was to standardise the group of trial examiners using the current rating scale, and then do multiple marking of a set of scripts. Nine raters – identified as representative of the 'universe' or worldwide population of raters for ESOL tests – rated a total of 25 writing performances.

A principal aim of the marking trial was to improve both the reliability and validity of the writing assessment process for SfL writing by refining the rating scale. It was hoped that the combined use of quantitative methodologies (application of criteria and scales to sample language performance) and qualitative methodologies (insightful and intuitive judgements derived from 'expert' participants) would inform the revision of the SfL writing rating scale.

Previous studies into the validation of writing rating scales have focused primarily on aspects such as administrative feasibility, face validity, training activities and scale validation (Shaw 2001, 2003). The trial described here is mainly concerned with scale validation and training activities.

Background to Skills for Life

The national strategy to tackle the literacy, language and numeracy needs of adults was launched by the government in March 2001 (see *Research Notes* 19). The strategy introduced, amongst other things, '*core curricula for literacy, numeracy and ESOL, to clarify the skills, knowledge and understanding that learners need in order to reach the national standards*' (Adult ESOL Core Curriculum 2001:1). The new ESOL core curriculum is based on the national standards for adult literacy developed by the Qualifications and Curriculum Authority (QCA) in 2000. The curriculum offers a framework for English language learning. It defines in considerable

detail the skills, knowledge and understanding that non-native English speakers need in order to demonstrate achievement of the national standards. It offers teachers of ESOL a reference instrument in a wide range of settings, which includes further and adult education, the workplace, programmes for the unemployed, prisons, community-based programmes, and family learning programmes.

The national standards comprise two parts: the standards themselves, which are the 'can do' statements, and the level descriptors, which describe in more detail what adults have to do to achieve the standards. Literacy covers the ability to speak, listen and respond; read and comprehend; and write to communicate.

The national standards for adult literacy and numeracy are specified at three levels: Entry Level, Level 1 and Level 2. Entry Level is further divided into three stages: Entry 1, Entry 2 and Entry 3. Entry Level is set out in this way to provide detailed descriptions of the early stages of learning in each skill. This sub-division also signals an alignment of the Entry stages with levels 1, 2 and 3 of the National Curriculum.

Writing and Skills for Life

The Adult ESOL Core Curriculum has been organised by level across the four skills of speaking, listening, reading and writing. The Adult Literacy and Adult ESOL core curricula both employ the overarching framework for teaching writing that is also used in the National Literacy Strategy for schools. According to the Adult ESOL Core Curriculum (2001:1), '*The complexities of the writing process are recognized by the model and the different levels on which fluent writers operate:*

- *Text level addresses the overall meaning of the text, the ability to write in different styles and formats*
- *Sentence level deals with grammar, sentence structure and punctuation*
- *Word level looks at the individual words themselves, their structure, spelling and meaning*'.

To develop understanding of the principles underpinning writing, the teacher may unpick different features at text, sentence or word level, but always with the final objective of producing an entire text.

In line with the Adult ESOL Core Curriculum, the Skills for Life mark scheme developed by Cambridge ESOL comprised three focuses, 'Text', 'Sentence' and 'Word' at all three Entry Levels. Each focus, in turn, comprised various assessment criteria.

In terms of levels of writing, the Adult ESOL Core Curriculum (2001:26) specifies the following:

- **At Entry Level 1**, Adults can be expected to write to communicate information to an intended audience.
- **At Entry Level 2**, Adults can be expected to write to communicate information with some awareness of the intended audience.
- **At Entry Level 3**, Adults can be expected to write to communicate information and opinions with some adaptation to the intended audience.

The mark scheme trial

In total, twelve people participated in the study: one Cambridge ESOL trainer (the officer responsible for the SfL Writing paper), nine examiners and two members from the Research and Validation Group. The examiner group consisted of a range of experienced EFL/EAP teachers and examiners. They had considerable experience of examining for Lower Main Suite, Upper Main Suite, BEC and IELTS. Table 1 summarises the examiners' background.

Table 1: Examiner Background

No. years experience as an EFL/EAP teacher			No. of years experience as a writing examiner		
Mean	**Max**	**Min**	**Mean**	**Max**	**Min**
15	31	4	16	30	2

Thirty scripts (10 Entry Level 1, 10 Entry Level 2 and 10 Entry Level 3) were made available for the trial. Every script was identified with a unique code according to its level and script number and subsequently distributed to examiners. Each Entry Level paper comprised three tasks so a total of 90 tasks were available for rating across all three levels. However, time constraints meant that only half of the Entry Level 3 scripts were marked by the examiners.

Examples of three Entry Level 3 tasks are given below together with what marks are awarded in each task.

Task 1: Completing a form; Marks awarded for 'Word' and 'Sentence'

You have just started an English language course at your local college. In the first lesson of your course, your teacher gives you a questionnaire to fill in.

Answer the questions.

ENGLISH AND YOU!

Name:

Age:

When did you start learning English?

How long have you been in the UK?

What language(s) do you speak apart from English?

How important are the following skills for you when you are learning English?

Reading
Writing
Listening
Speaking

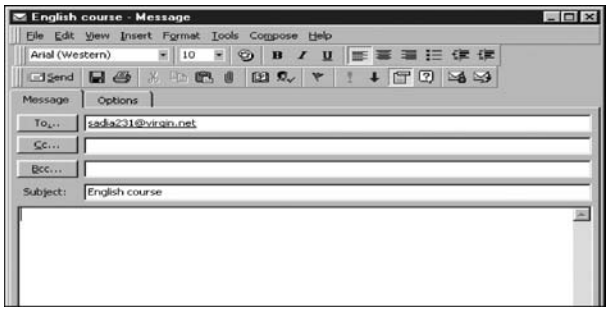
Please give your reasons for your choice(s).
(Write 20–30 words. Write in sentences.)

What do you like about learning English?
(Write about 20–30 words. Write in sentences.)

Task 2: Responding to an e-mail; Marks awarded for 'Sentence' and 'Text'

You decide to write an e-mail to your friend, Sadia, telling her about your teacher, describing the lessons and suggesting that she joins the course.

Write about 60 words.



Task 3: Writing a report; Marks awarded for 'Word', 'Sentence' and 'Text'

As part of your new course, your class is preparing a project about the college. Your teacher has asked you to write a report about ONE of the facilities in the college, for example the library, the sports centre or the cafeteria.

Write a report describing the facility you have chosen, explaining what you like about it and suggesting ways to make it better.

You do not need to use a report format.

Write about 120–150 words.

The group of trial examiners were standardised using the current rating scale, and then undertook multiple marking of a set of scripts. The first part of the session involved an overview of the trial and examiner induction during which the examiners were trained and briefly standardised with the SfL Writing assessment criteria and band level descriptors. They then marked Entry Level 1 scripts first, followed by Entry Level 2 and then Entry Level 3. After the marking was finished, the examiners completed a questionnaire focusing on different aspects of the mark scheme. The session ended with an open-ended discussion of the examiners' experience using the SfL mark schemes.

Data analysis

The data analysis encompassed both quantitative and qualitative methods. The quantitative methodologies included correlational analyses, computation of examiner strength of agreement indices and inter-rater reliabilities (across each of the three Entry Levels) and analysis-of-variance. The qualitative methodologies included individual verbal protocols collected during actual marking, a questionnaire and an open-ended plenary session. The protocols were used to explore the cognitive process that the examiners undertook to mark and identify issues directly related to the mark schemes. The retrospective data captured by the examiner questionnaire was used to supplement the plenary session – essentially, a semi-structured focused discussion group designed to enable the examiners to share their views with each other and with the Cambridge ESOL team. Examiners further recorded their own observations during marking in order to corroborate the verbal protocols by recording issues related to the understanding, use and application of the mark schemes, and to act as a 'back up' in case the verbal protocols were unfruitful.

Statistical analysis

Results of the quantitative analyses provided evidence of the validity and reliability of the rating scales. More specifically, descriptive statistics and analysis-of-variance (ANOVA) indicated that the examiners were generally homogeneous in the marks awarded. The ANOVA revealed no significant differences between the raters at:

Entry Level 1 – $F(6, 49) = .296$, $p > .05$;

Entry Level 2 – $F(6, 49) = 1.479$, $p = > .05$, and

Entry Level 3 – $F(6, 28) = .411$, $p = > .05$.

In addition to the ANOVA findings, mean examiner intercorrelations (Spearman's rho) were found to be consistently high – of the order of 0.9, ranging from .802 to .976 for Entry Level 1; .663 and .898 for Entry Level 2 (with three lower correlations restricted to three specific raters); and .667 and 1.000 for Entry Level 3. On this basis, we could be confident that the examiners shared approximately 81% similar information with only 19% of information unaccounted for.

Statistical significance tests further indicated that the strength of the Spearman correlational findings was such that there was evidence of a strong relationship between the examiners. Kendall's Coefficient of Concordance suggested that there was a significant relationship in how the raters viewed assessment criteria for each of the three Writing assessment focuses (Word, Sentence, Text). More specifically, the results indicated that we can be 95% confident that the distribution of ranking was not necessarily random. Examiners were clearly distinguishing between the three assessment criteria.

Strength of Agreement indices revealed levels of agreement between the evaluations of SfL trial raters when rating the same sets of scripts. Strength of Agreement testing indicated only a low level of rater agreement regarding category membership of each script. Interestingly, inter-rater reliabilities were high (ranging from .77 to .79). Computation of an inter-rater index is related to, and dependent upon, correlation. Whilst correlation is an indication of rater consistency it is also a measure of a rater's ability to rank order their marking of a set of scripts. Clearly, trial raters were in general agreement on the rank ordering of the scripts although they were in less agreement regarding the absolute mark assigned to those scripts. Kappa has the advantage that the agreement matrix generated for the calculation of the Kappa statistic shows clearly the agreement and disagreement amongst raters. The potential value of this is great for the training of SfL examiners. The technique can point out where disagreements most often occur. Future training of examiners can, therefore, be targeted at achieving better consensus in the categories where the greatest discrepancies lie. The test can also indicate which scripts are most effective for future training/certification sets.

Supplementary findings from complementary qualitative studies revealed that examiners were, for the most part, enthusiastic about the marking experience, considering it to be both rewarding and positive. Despite raising a number of concerns, examiners were favourably disposed to the SfL mark schemes. In general, the trial demonstrated the potential of revising the mark scheme into a more user-friendly version and the raters were favourably disposed towards the new rating approach.

Examiner questionnaire

The survey of Cambridge ESOL writing examiners, albeit small, aimed to deduce:

- how they felt about the writing assessment procedures
- how, as individuals, they rated written responses to each Entry Level for each task in relation to the SfL assessment criteria and band level descriptors
- how they felt about the assessment criteria and band descriptors
- their general impressions and opinions of the mark scheme thus far and its implications for training.

The questionnaire consisted of five sections:

1) Examiner Background Information
2) Using the Skills for Life criteria and band descriptors – General Rating Issues
3) Using the Skills for Life criteria and band descriptors – Word Focus
4) Using the Skills for Life criteria and band descriptors – Sentence Focus
5) Using the Skills for Life criteria and band descriptors – Text Focus.

The questionnaire responses indicated that in general the raters felt they understood the descriptors well and could use the mark scheme accurately. Their responses – regarding what they found most/least difficult to assess and what they paid most/least attention to – provided some insights into future examiner training. For example, in terms of what examiners found most difficult to assess, the examiners noted that at the word level, they found handwriting, vocabulary and digit formation most difficult; at the sentence level they found grammatical range, complex sentence structures used with lower-level vocabulary and word order most difficult to assess; at the text level they found register, task fulfilment and coherence most difficult to assess. Such feedback provided useful insights into the way the raters approached and used the mark scheme. They were later incorporated as specific issues to be addressed in examiner training.

In terms of productivity, examiners were concerned that – at least initially – their overall rate of marking would be affected: many mark schemes eliciting longer reading and processing time and ultimately lengthening the rating process. Most of the examiners' specific comments centred around the content of the mark scheme and the descriptors themselves. For example, the layout and length of the descriptors were generally felt to need revision: '*There are problems assimilating the band descriptors because they are lengthy and dense. To make them easier to use, they could be organised in bullet format.*' The use of exemplification in the mark scheme brought forward opposing views, from the barely complimentary to the highly favourable: '*The amount of information, i.e. examples, included is very supportive.*'

The examiners also made specific suggestions for 'word', 'sentence' and 'text' level descriptors. At the 'word' level, they made comments on issues connected to spelling, supplying an address in a form, giving a postcode, use of upper/lowercase letters, formation of words, giving a signature, and assessing handwriting. At the 'sentence' level, they commented on underlength responses and incomplete sentences. At the 'text' level, they commented on the need to include more guidance on irrelevant content, use of appropriate formulae for a specific genre or register and the issue of task achievement.

The examiners also made comments on issues they felt were not explicit enough in the draft mark scheme and suggested including additional descriptors for the following issues: penalising for inadequate length, missing content points, misunderstanding of tasks, range of vocabulary being prominent in the descriptors.

The verbal reports and extended written feedback raised some specific questions related to the layout of the mark scheme, question paper and optically read marksheet, the rate of marking, the number of available marks and some general and specific descriptor issues. In terms of the mark scheme layout, the examiners noted that the overall presentation of mark schemes should be improved and gave specific suggestions on how to make it more 'user-friendly'.

All the examiner feedback was used in the subsequent revision of the Skills for Life mark scheme. Based on some of the comments, it was felt that the layout of the mark scheme could benefit from explicit mention of the assessment criteria underlying the 'Word', 'Sentence' and 'Text' focuses (see Table 2 below). Structuring the mark scheme in this way would lead to more clarity and transparency, which would potentially improve rater accuracy and reliability.

Table 2: Skills for Life Writing assessment criteria

WORD FOCUS
Handwriting
Spelling
Upper/lower case letters
Vocabulary (in some tasks)
SENTENCE FOCUS
Grammatical range and accuracy
Word order
Mechanical features (including punctuation and the pronoun 'I')
TEXT FOCUS
Content
Organisation
Register or stylistic features (depending on the task and level)

Conclusion

It is essential to the success of the SfL Writing test that effort is given to the validation of the rating scale prior to its widespread use. To this end a range of qualitative and quantitative methods for the establishment of validity, reliability, impact and practicality (VRIP) has been undertaken.

Although this phase of the work was designed to be exploratory, these results have enabled Cambridge ESOL to revise and improve the existing Skills for Life Writing mark scheme using empirically based findings. It has also allowed the researchers to develop a more focused perspective for any subsequent review of the mark scheme as well as providing insightful observations into the way examiners mark. The trial has engendered confidence in the mark schemes which will be used after the live release of the Skills for Life tests in March 2005.

References and further reading

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C and Milanovic, M (Eds), *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Cambridge: UCLES/Cambridge University Press, 57–120.

Shaw, S (2001) The effect of standardisation training on rater judgement and inter-rater reliability for the revised CPE Writing paper 2, Cambridge ESOL: UCLES internal report no. 290.

—(2003) IELTS Writing Assessment Revision Project (Phase 3): Validating the revised scale. A qualitative Analysis, Cambridge ESOL: UCLES internal report no. 514.

Applying lexical statistics to the IELTS speaking test

JOHN READ, VICTORIA UNIVERSITY OF WELLINGTON

Introduction

As discussed in previous issues of *Research Notes* (Ball 2001, 2002), the creation of corpora consisting of samples of written and spoken production by examination candidates promises to be a fruitful basis for validation research at Cambridge ESOL. Corpus analysis lends itself particularly well to the study of vocabulary use because word forms can be readily identified and counted by a computer program. This has greatly enhanced fields such as lexicography, ELT curriculum and materials design, and second language vocabulary research by improving the quality of information about the relative frequency of words and their meanings in everyday usage. In a testing context, lexical analysis of a corpus can give a useful profile of how the vocabulary use of candidates varies at different levels of proficiency and also the extent to which learners differ from native speakers in the size and range of their vocabularies.

There is a long tradition of lexical analysis of written texts by means of various simple statistics, which I reviewed in my book *Assessing Vocabulary* (Read 2000). In addition, child language researchers and educationalists use vocabulary measures to track the increasing linguistic and cognitive maturity of young people as they go through childhood and adolescence to adult life. However, relatively little is known about the spoken vocabulary of either native speakers or second language learners, and it is only recently that researchers are beginning to fill this gap through investigations of the large spoken corpora that are now available (see Adolphs and Schmitt 2003, McCarthy and Carter 2003).

I had the opportunity to explore this area of spoken vocabulary use in a testing context through a project funded by IELTS Australia as part of the IELTS Funded Research Program. In collaboration with my colleague Paul Nation, I conducted a study of vocabulary use by candidates in the revised version of the IELTS Speaking Module, which was introduced in July 2001. The revision brought several changes to the test, two of which had interesting implications from a vocabulary perspective:

- The test moved from a five-part to a three-part structure, consisting of an interview; a long turn, in which the candidate speaks at length on a specified topic; and an extended discussion based on the same topic. Thus, for individual candidates a single topic tends to dominate the test.
- Instead of the previous single holistic scale, the examiner now rates the candidate's performance on four analytic scales. Three of them are fluency and coherence; grammatical range and accuracy; and pronunciation; and the fourth is lexical resource. This means that examiners are required to pay some specific attention to the quality of the candidates' vocabulary use.

Thus, we set out to find what might be revealed by applying various lexical statistics to a mini-corpus of performances by IELTS candidates in their speaking test.

Creating the corpus

For this study, we obtained a total of 88 recordings of IELTS speaking tests selected from a much larger set of tapes sent to Cambridge ESOL from test centres in various parts of the world as part of the routine monitoring procedures for the test. There were two main criteria for selecting the tapes. First, the choice was restricted to just four of the available Part 2 topics, in order to control the topic effect on vocabulary use; and secondly, we wanted an even number of candidates at each of the proficiency levels from Band 4 to Band 8, although in practice there were somewhat fewer at those two band scores than at Bands 5, 6 and 7. The final sample of tapes were recorded at 21 test centres in 14 countries around the world.

Initially the tapes were fully transcribed, but in order to apply various lexical statistics to each candidate's speech, it was necessary to create a new text with the examiner's contributions deleted and some editing of the candidate's spoken production. For example, hesitation utterances such as *um*, *mm* and *er* were removed, as were items such as *mhm* and *OK*, in cases where the candidate was just acknowledging what the examiner was saying. In addition, false starts and repetitions of words or short phrases (eg, *it's the the standard* or *in the sense in the sense that…*) were taken out. Another problem was that some sections of certain tapes had been difficult to transcribe either because of the poor quality of the recording or unclear pronunciation by the candidate. Thus, as compared to written texts, there are real challenges in determining what the basic lexical elements of an oral text are, and careful editing is needed to ensure that the statistical results will be as reliable as possible.

Once the edited texts had been prepared, various lexical analyses could be applied. There are now several software programs which are available for this purpose. One widely used package, *WordSmith Tools* (Scott 1998), has a variety of applications in corpus analysis, but the basic tools are word frequency counts, the display of individual words in context (or concordancing), and a 'keyword' analysis, which compares the relative frequency of words in particular text(s) with their frequency in a more general reference corpus. Other more specialised programs produced by researchers on second language vocabulary learning are Range (Nation and Heatley 1996), P_Lex (Meara and Bell 2001) and D_Tools (Meara and Miralpeix 2004). The applications of these analyses in the present study are discussed below.

Results

Lexical output

The first step was to look simply at how many words were produced by candidates at the five band levels represented in our 88 texts. Table 1 shows lexical output both in terms of **tokens** (the total number of words used) and **types** (the number of *different* words used). Since there were varying numbers of candidates at the five band levels, the means in the right-hand columns of the table give a more accurate indication than the raw totals, and they show that – as you might expect – more proficient candidates produced more vocabulary than less proficient ones. However, the standard deviations also show wide variation in the number of words spoken by candidates *within* each band level. By itself, then, lexical output has limited value as a measure of the quality of the candidate's speaking performance.

Lexical variation

The counting of types and tokens provides the basis for another vocabulary measure, which is called lexical variation or diversity. Traditionally, this has been calculated as the Type-Token Ratio (TTR), the ratio of the types to the tokens. A high TTR shows that the language user has produced a large proportion of different words (types), whereas someone with a low ratio tends to make repeated use of a smaller number of types. One long-recognised problem with the TTR, though, is that the ratios are affected by the length of the texts. This made it an unsuitable measure to use with our speaking tests, because there was so much diversity in the number of tokens produced by the candidates.

Table 1: Lexical output of IELTS candidates by band score level

	TOTALS		MEANS (SD)	
	Tokens	**Types**	**Tokens**	**Types**
BAND 8 (n=15)	22,366	2374	1491.0 (565.9)	408.1 (106.0)
BAND 7 (n=19)	21,865	2191	1150.7 (186.7)	334.6 (46.0)
BAND 6 (n=19)	18,493	1795	937.3 (261.4)	276.7 (48.2)
BAND 5 (n=21)	15,989	1553	761.4 (146.7)	234.2 (35.5)
BAND 4 (n=14)	6,931	996	475.8 (216.9)	166.6 (48.6)

To overcome this limitation, Malvern, Richards and their associates (Malvern and Richards 2002; Durán, Malvern, Richards and Chipere 2004) have recently devised a more sophisticated method of measuring lexical variation, which involves taking multiple samples of words from the text and employing curve-fitting techniques to calculate a value called D (or Diversity), which ranges from 0 to 90. In this study, Meara and Miralpeix's (2004) D_Tools program was used to make the calculation.

The D values for the texts in our corpus are presented in Table 2. The pattern of the findings for lexical variation is somewhat similar to those for lexical output. The mean values for D decline as we go down the band score scale, but again the standard deviations show a large dispersion in the values at each band level, and particularly at Bands 7 and 6. Thus, as a general principle, more proficient candidates use a wider range of vocabulary than less proficient ones, but D by itself cannot reliably distinguish candidates by band score.

Table 2: Summary output from the D_Tools Program, by band score level

	D (Lexical Diversity)			
	Mean	**SD**	**Maximum**	**Minimum**
BAND 8 (n=11)*	79.0	4.9	87.5	72.0
BAND 7 (n=17)*	71.8	18.2	89.5	61.2
BAND 6 (n=18)*	67.2	16.0	81.4	57.0
BAND 5 (n=21)	63.4	11.3	86.7	39.5
BAND 4 (n=14)	60.7	11.4	76.1	37.5

* Seven candidates with abnormal D values were excluded

Lexical sophistication

A third kind of vocabulary measure is known as lexical sophistication. This is usually conceived as being the proportion of relatively unusual or low-frequency words in a text. For this purpose, we can use Nation and Heatley's (1996) Range program, which classifies the vocabulary used in a text into four categories, represented by the columns in Table 3. The first two columns record the percentage of high-frequency words, i.e., those among the first and second thousand most frequent words in English, based on West (1953). The third column covers words in the Academic Word List (Coxhead 2000), a specialised inventory of 570 words occurring with high frequency in academic texts. The remaining words, not in the three preceding lists, are represented by the fourth column. This latter category includes technical terms, proper names, colloquial terms and other lower-frequency words.

Table 3: The relative percentages of high- and low-frequency words used by candidates at different band score levels

	TYPES				
	List One	**List Two**	**List Three**	**Not in Lists**	**Total**
BAND 8 (n=15)	1270 53.7%	347 14.7%	243 10.3%	504 21.3%	2364 100%
BAND 7 (n=19)	1190 54.6%	329 15.1%	205 9.4%	455 20.9%	2179 100%
BAND 6 (n=19)	1060 59.5%	266 14.9%	179 10.0%	277 15.5%	1782 100%
BAND 5 (n=21)	958 62.1%	222 14.4%	119 7.7%	243 15.8%	1542 100%
BAND 4 (n=14)	677 68.5%	132 13.3%	58 5.9%	122 12.3%	989 100%

List One: 1st 1000 words of the GSL (West 1953)
List Two: 2nd 1000 words of the GSL
List Three: Academic Word List (Coxhead 2000)
Not in the Lists: Not occurring in any of the above lists

In broad terms, Table 3 shows the type of pattern that we might expect. Assuming that candidates at Band 4 have relatively limited vocabulary knowledge, we can see that Lists One and Two cover more than 80% of the words they used in the test, whereas this high-frequency vocabulary accounts for less than 70% of the words produced by Band 7 and 8 candidates. Conversely, the percentages of academic and lower frequency words decline as we go down the table from Band 8 to Band 4. In that sense, then, the vocabulary use of higher proficiency candidates was somewhat more sophisticated than that of those at low band levels, but on the other hand, all candidates used a high percentage of high-frequency words and the Range analysis offers a rather crude measure of the quality of their lexical expression.

Another perspective on the lexical sophistication of the speaking texts is provided by Meara and Bell's (2001) P-Lex program, which produces a summary measure, lambda, based on this same distinction between high- and low-frequency vocabulary use in individual texts. A low value of lambda shows that the text contains mostly high-frequency words, whereas a higher value is intended to indicate more sophisticated vocabulary use.

Table 4: Summary output from the P-Lex Program, by band score level

	LAMBDA			
	Mean	**SD**	**Maximum**	**Minimum**
BAND 8 (n=15)	1.10	0.22	1.50	0.77
BAND 7 (n=19)	1.05	0.26	1.49	0.60
BAND 6 (n=19)	0.89	0.17	1.17	0.55
BAND 5 (n=21)	0.88	0.24	1.38	0.33
BAND 4 (n=14)	0.83	0.33	1.48	0.40

As shown in Table 4, the mean values of lambda show the expected decline from Band 8 to 4, confirming the pattern in Table 3 that higher proficiency candidates used a greater proportion of lower-frequency vocabulary in their speech. However, the standard deviations and the range figures also demonstrate what was seen in the earlier tables: except to some degree at Band 6, there was a great deal of variation within band score levels.

Topic-specific vocabulary

In this one further analysis, we turn our attention from variation in vocabulary use by band level to variation according to topic. As previously noted, each candidate is assigned a topic to speak about for the long turn in Part 2 of the speaking test and this same topic is the starting point for the discussion in Part 3. For our corpus we included four topics (or 'tasks', as they are called in the IELTS programme), which can be identified briefly as follows:

Task 1: A favourite restaurant

Task 2: A favourite book

Task 3: Learning English

Task 4: An admired person

WordSmith Tools offers two kinds of analysis that are helpful in identifying topic-specific words. One is 'WordList', which produces frequency lists of the words that occur in the 22 or so texts of candidates who responded to each of these four tasks. The other is 'KeyWords', which compares the frequency of words in a particular set of texts with their frequency in a broader reference corpus, in order to identify – in this case – the words that were characteristic of IELTS speaking tests based on a given Part 2 task. For our KeyWords analysis, we used the other three tasks collectively as the reference for each set of texts. For instance, in order to locate the key words in Task 1, we compared the word frequencies in Task 1 texts with the frequencies of words in Tasks 2, 3 and 4 taken together. WordSmith Tools generates a 'keyness' statistic, which is the result of a chi-squared test of the significance of the difference between the frequency of the word in the two sets of texts. Table 5 lists the words with the fifteen highest keyness values for each of the four tasks.

Table 5: Results of the KeyWord analysis for the four Part 2 tasks

TASK 1		TASK 2		TASK 3		TASK 4	
Favourite Restaurant (n=23)		**Favourite Book** (n=22)		**Learning English** (n=21)		**Admired Person** (n=21)	
food	463.1	read	342.8	English	713.1	he	346.5
restaurant	327.8	books	309.2	language	233.6	famous	270.4
fast	184.0	book	358.9	learn	251.1	people	115.2
eat	104.8	reading	102.2	speak	99.4	him	110.6
foods	90.0	story	66.4	learning	76.8	person	76.0
eating	86.7	children	57.2	languages	74.7	his	60.2
go	76.1	internet	38.4	school	72.4	public	53.0
cook	74.3	television	38.4	class	69.7	admire	51.5
like	58.8	girl	36.8	grammar	62.2	who	50.6
home	57.7	men	36.8	communicate	56.2	known	48.5
traditional	52.0	writer	35.1	foreign	52.1	media	45.7
restaurants	47.0	boy	29.7	started	40.5	become	42.0
dishes	45.3	this	28.6	words	37.7	she	39.0
cooking	45.3	hear	28.5	speaking	34.9	chairman	24.2
nice	42.2	women	27.4	teacher	33.8	president	24.2

Clearly the KeyWord analysis shows a strong task effect on the IELTS speaking test. The words in the lists represent in a sense the default vocabulary for each topic: the mostly high-frequency words one would expect learners to use in talking about the topic. As such these words will almost certainly not be salient for the examiners in rating the learners' lexical resource, except perhaps in the case of low-proficiency candidates who exhibit uncertain mastery of even this basic vocabulary.

One other interesting observation to come out of the WordSmith analyses is some variation among the four tasks. Tasks 1 and 3 tended to produce a smaller number of keywords, apparently because candidates talked about these topics using quite similar vocabulary. For example, for Task 3 the focus was generally on the formal study at English at school and so words like *learn*, *school*, *class* and *teacher* figured prominently. On the other hand, in the case of Task 2, the books that candidates chose to talk about were diverse in content and so the keywords tended to be those which were frequently used in the Part 3 discussion about the reading habits of men, women and children in the candidate's country, rather than words used to describe what the book was about in Part 2 of the test.

Conclusion

The application of these various statistical analyses to the vocabulary use of candidates in the IELTS speaking test has produced some interesting – but not really unexpected – findings. In general terms, the mean values of the statistics distinguish candidates at different band scores. Highly proficient candidates at Bands 7 and 8 produce more speech and use a wider range of lower frequency vocabulary, certainly as compared to candidates who are at Bands 4 and 5. At the same time, there is considerable variance within band levels for all of these measures.

It is not surprising, then, if examiners have some difficulty in reliably rating the Lexical Resource of IELTS candidates, particularly in differentiating performance at adjacent band score levels. Further research is needed to follow up the present study by obtaining verbal reports from examiners as they rate an IELTS candidate's performance to see which features of vocabulary use are salient to them in the test situation and might influence the band score they record for Lexical Resource. In fact, Annie Brown has taken this approach in her project in Round 9 of the IELTS Funded Research Program to investigate the rating process in the speaking test.

One limitation of the statistics reported here is that they deal with individual word forms, rather than larger lexical items such as compound nouns, phrasal verbs, colloquial expressions, idioms and so on. If we extend the concept of a learner's lexical resource beyond individual words, we may well find that sophistication in vocabulary use in the speaking test involves fluent use of idiomatic expressions composed of highly frequent words, perhaps more so than a mastery of low-frequency academic or technical words. In our project, we did some qualitative analyses of the full transcripts along these lines and found some evidence to support this proposition. It is beyond the scope of the present brief report to present the findings here, but suffice it to note in conclusion that, in order to gain a fuller understanding of the lexical component of candidates' performance in the IELTS speaking test, it is necessary to complement the statistics on the occurrence of individual words with qualitative analyses of vocabulary use in a more general sense.

References and further reading

Adolphs, S and Schmitt, N (2003) Lexical coverage of spoken discourse, *Applied Linguistics*, 24/4, 425–438.

Ball, F (2001) Using corpora in language testing, *Research Notes*, 6, 6–8.

—(2002) Developing wordlists for BEC, *Research Notes*, 8, 10–13.

Coxhead, A (2000) A new academic word list, *TESOL Quarterly*, 34/2, 213–238.

Durán, P, Malvern, D, Richards, B and Chipere, N (2004) Developmental trends in lexical diversity, *Applied Linguistics*, 25/2, 220–242.

Malvern, D and Richards, B (2002) Investigating accommodation in language proficiency interviews using a new measure of lexical diversity, *Language Testing*, 19/1, 85–104.

McCarthy, M and Carter, R (2003) What constitutes a basic spoken vocabulary? *Research Notes*, 8, 5–7.

Meara, P and Bell, H (2001) P_Lex A simple and effective way of describing the lexical characteristics of short L2 texts, *Prospect*, 16/1, 5–24.

Meara, P and Miralpeix, I (2004) *D_Tools*, Swansea: Lognostics (Centre for Applied Language Studies, University of Wales Swansea).

Nation, P and Heatley, A (1996) *Range* [Computer software], Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Read, J (2000) *Assessing Vocabulary*, Cambridge: Cambridge University Press.

Scott, M (1998) *WordSmith Tools*, Version 3.0, [Computer software], Oxford: Oxford University Press.

West, M (1953) *A General Service List of English Words*, London: Longmans Green.

Upper Main Suite speaking assessment: towards an understanding of assessment criteria and oral examiner behaviour

EVELINA GALACZI, RESEARCH AND VALIDATION GROUP

Introduction

Candidate performance on Upper Main Suite (UMS) examinations is routinely monitored at Cambridge ESOL with a focus on issues central to speaking assessment such as scoring criteria and behaviour of Oral Examiners. The findings and insights gained from such investigations contribute to an ongoing research agenda which seeks to provide evidence for the validity of the Cambridge ESOL examinations.

This article reports on a study of the Speaking tests for Upper Main Suite examinations (FCE, CAE and CPE) in 2003. More specifically, the focus is on:

- Overall candidate performance on the scoring criteria
- Levels of agreement between the Interlocutor and Assessor marks
- Correlations between the different assessment criteria.

Background to the Upper Main Suite Speaking Test

The standard Cambridge approach to speaking assessment involves a face-to-face paired format, i.e., two candidates and two oral examiners. One examiner acts solely as Assessor and does not join in the interaction; the second examiner – the Interlocutor – manages the Speaking test by asking questions and providing cues for the test takers.

All three levels in the Upper Main Suite of examinations share common features in terms of test format, test materials, assessment criteria and procedures and the roles of Oral Examiners. The distinguishing features of the Upper Main Suite speaking test are:

- *A paired test format* – the standard test format is 2:2, i.e., two candidates and two examiners. A trio format is used in cases where there is an uneven number of candidates in an examining session and the last candidate joins the final pair of candidates to form a group of three.
- *A multi-task test* – at FCE and CAE level, the Speaking test comprises four tasks and at CPE level three tasks. Each task elicits a different type of language and uses a different interaction pattern, e.g., conversation between the interlocutor and candidate; individual long turn from each candidate; two-way interaction between the candidates; three-way interaction between the candidates and interlocutor.
- *The use of an Interlocutor frame* – the interlocutor frame is a script for the Interlocutor which specifies the language to be used when introducing and managing a task and the materials to be given to the test takers. The interlocutor frame is used for the purposes of standardisation, since by adhering to the script, examiners ensure that all candidates are treated equally.
- *Interlocutor/Assessor assessment* – assessment is provided by both the Interlocutor and Assessor. Candidate output is assessed based on performance on all four tasks. The Interlocutor provides a global assessment, while the Assessor provides analytic assessment on a range of criteria. Both assessments are based on a nine-band mark scheme, where 1.0 is the lowest score, 5.0 the highest with intervals of half scores. Each examiner views the performance from a different perspective and arrives at independent marks using a different set of criteria.
- *Common scale for speaking* – the assessment procedures and criteria are based on the Cambridge ESOL Common Scale for Speaking (ffrench 2003), in which the range of abilities of the candidates are assessed against a set of analytic assessment criteria and a global mark. The FCE and CAE assessment

Table 1: Overall candidate performance by marking criteria

Assessment Criteria	FCE		CAE		CPE	
	Mean (Out of 5)	SD	Mean (Out of 5)	SD	Mean (Out of 5)	SD
Grammar and Vocabulary	3.53	0.70	3.33	0.74	-	-
Grammatical Resource	-	-	-	-	3.26	0.80
Lexical Resource	-	-	-	-	3.39	0.79
Discourse Management	3.66	0.72	3.50	0.73	3.48	0.79
Pronunciation	3.72	0.63	3.60	0.65	3.56	0.72
Interactive Communication	3.78	0.70	3.63	0.70	3.61	0.74
Global Achievement	3.68	0.70	3.49	0.71	3.46	0.77

The data given comprise the average score for all respective 2003 sessions

criteria comprise four analytical scales (Grammar and Vocabulary, Discourse Management, Pronunciation and Interactive Communication). In the case of CPE the Grammar and Vocabulary component is sub-divided into two criteria (Grammatical Resource and Lexical Resource).[1] The analytic assessment criteria are applied to the candidate's performance across the whole test by the Assessor. In addition to the analytic scale, a separate scale – the Global Achievement Scale – is used by the Interlocutor to assess the candidate's overall effectiveness in tackling the tasks. The Global Achievement mark is not the average of the analytical marks, but a global mark reflecting an assessment from a different, more holistic, perspective.

Findings

Scoring criteria

• *What is the overall performance of candidates by marking criteria?*

An analysis of the mean scores for the analytic scales (see Table 1) revealed that across all three examinations candidates achieved the lowest mean scores on either 'Grammar and Vocabulary' (for FCE and CAE) or 'Grammatical Resource' and 'Lexical Resource' (for CPE). The highest mean score was observed with 'Interactive Communication'.

Based on the above analysis it is difficult to say with certainty why some scoring criteria behaved differently than others in terms of overall performance. The lowest means for 'Grammar and Vocabulary' and 'Grammatical Resource'/'Lexical Resource' could be a result of the fact that these two elements in the mark scheme were more noticeable and measurable and as a result marked more 'harshly' by oral examiners.

In the case of the highest mean for 'Interactive Communication', we could speculate that the construct of 'Interactive Communication', in comparison with other constructs (e.g. 'Grammar and Vocabulary' or 'Pronunciation') is positively defined and marked. In other words, examiners reward when they hear good examples, rather than mark negatively when they hear mistakes.

The descriptive statistics presented in Table 1 further indicate that the 'Pronunciation' criterion consistently revealed the lowest standard deviation for each of the three levels, suggesting that this criterion may be less discriminating than the others, with the raters using a 'shorter' scale than they did for the other elements. The issue of the lower variance of the Pronunciation scores has been investigated internally at Cambridge ESOL by Green (2004), who focused on the performance of the Pronunciation scale in relatively homogeneous and heterogeneous cohorts with respect to L1. A possible explanation for the seemingly 'shorter' nature of the Pronunciation scale was that candidates sharing the same L1 are also relatively homogeneous in their pronunciation and that the large numbers of candidates in certain cohorts are affecting the overall distribution of scores. The findings confirmed that heterogeneous cohorts display greater variance in the Pronunciation scale, and provided some evidence that the degree of variance of this scale is associated with L1 features. In other words, in a multi-lingual test context, the Pronunciation scale was at least as widely used as the other scales.

Inter-rating agreement: Interlocutor and Assessor

• *What is the correlation between the scores given by the Interlocutor and the Assessor for each exam?*

This issue was addressed by estimating the Pearson correlation between the Global Achievement score given by the Interlocutor and the combined analytical scores given by the Assessor. It needs to be noted that the correlation between the ratings awarded by the Interlocutor and Assessor is not necessarily the same as inter-rater reliability. Inter-rater reliability refers to the agreement between different examiners using *identical* scoring procedures. The Cambridge approach to scoring speaking tests is based on the principle of complementary perspectives where two examiners rate a candidate's performance from different viewpoints. One is a

1. For a detailed description of the assessment criteria, see www.CambridgeESOL.org/exams/

participant who assigns a global rating while interacting with the candidate (the Interlocutor); the second is an observer who does not interact with the candidates and assigns a set of analytical scores (the Assessor). It was therefore deemed important to refer to this level of agreement as *inter-rating* agreement.

A Pearson correlation of 1.00 implies perfect agreement between the two examiners. A complete agreement of this kind could only be achieved in theory when one single perspective is privileged and all competing perspectives are dismissed. In the Cambridge ESOL Main Suite examinations, the Interlocutor and Assessor arrive at their marks through different processes, taking different roles in the interaction. It is not expected, nor is it desirable, that they should be in complete agreement on the scores to be awarded. In general, a Pearson correlation between .65 and .85 is deemed acceptable in the Cambridge ESOL Main Suite examinations.

The inter-rating correlations between the Global Achievement mark (given by the Interlocutor) and the total of the four analytical marks (given by the Assessor) are presented in Table 2.

Table 2: Inter-rating agreement between the Interlocutor and Assessor

Examination	Range of inter-rating correlations for all 2003 sessions	Average
CPE	Between .793 and .872	.818
CAE	Between .791 and .798	.795
FCE	Between .789 and .846	.811

The findings in Table 2 show an acceptably high level of agreement between the marks given by the Assessor and those awarded by the Interlocutor. In addition, they indicate that while the two raters provided independent ratings of candidate performance, they were also in general agreement. The high level of agreement is evidence of the effectiveness of training, co-ordination and monitoring of Oral Examiners.

Inter-rating agreement: global mark and analytical components

- *What is the correlation between the global mark and each analytical component of the mark scheme for each exam?*

This issue was addressed in order to gain insights into the level of overlap between the different constructs underlying the UMS mark scheme. While it does not inform the operational requirements of the test, it is seen as providing potentially useful information regarding the construct validity of the marking components.

In general, in all three exams the highest correlations were observed between:

- 'Discourse Management' and 'Grammar and Vocabulary' (or 'Grammatical Resource'/'Lexical Resource' in CPE): in the .84 and .89 range
- 'Discourse Management' and 'Interactive Communication': in the .81 and .88 range.

These findings indicate that the construct of 'Discourse Management' overlapped the most with the constructs of 'Grammar and Vocabulary' and 'Interactive Communication'. In other words, having good discourse management control also meant good lexico-grammatical and turn-taking control. This overlap is not altogether surprising and could be illustrated with cohesive devices, for instance. Cohesive devices such as 'and', 'because', in addition', 'then' play a role in discourse management, vocabulary, and turn-taking.

In terms of the lowest correlations, 'Pronunciation' consistently correlated the lowest with the other mark scheme criteria (in the .61 and .71 range). These findings indicate that, not surprisingly, the construct of 'Pronunciation' seems distinctively different from the other analytic criteria. In other words, having good pronunciation did not necessarily entail good lexico-grammatical, discourse or conversation management control.

It is difficult to be certain what factors lie behind these two trends of low/high correlations between some of the mark scheme elements. A likely possibility could be the definition of the theoretical constructs in the mark scheme, or the nature of the constructs themselves. Given that for the purposes of assessment Cambridge ESOL defines the construct of speaking ability in terms of several components, it is reassuring to find evidence that some components, e.g. 'Pronunciation', emerge as distinct and relatively autonomous. At the same time, common sense indicates that there is a higher-level, overarching construct of language ability and the high correlations between some mark scheme components support that notion as well. This finding could be used as a starting point for investigations of the construct validity of the mark scheme. A future research endeavour which goes beyond simple Pearson correlations and focuses on the relationship between the constructs underlying the scoring rubric with more sophisticated research tools would reveal valuable insights in this area.

Another possible cause behind the correlations between the different mark scheme components could be the way the raters were interpreting the scale. Rater training and standardisation is a key issue for Cambridge ESOL and is the focus of ongoing research. More specifically, a currently ongoing programme of studies which focuses on rater strategies in the Cambridge speaking tests will allow us to be more certain of interpreting these findings in the light of raters' use of the mark scheme.

Conclusion

This study forms part of an ongoing validation programme for the UMS speaking tests, which in turn is part of Cambridge ESOL's model of test development which incorporates research and validation as essential components (Saville 2003). On the one hand, the present findings help to inform ongoing issues in speaking test research and support various aspects of the Cambridge speaking tests; on the other hand, they suggest future research possibilities or feed into studies already being carried out. Overall, these findings enable us to gain a greater understanding of the performance of speaking test candidates and their examiners and such deeper insights can contribute to a wider research agenda for the assessment of speaking, both for our Main Suite, general English qualifications and for other qualifications such as Skills for Life or IELTS described elsewhere in this issue.

References and further reading

ffrench, A (2003) The development of a set of assessment criteria for Speaking Tests, *Research Notes* 13, 8–16.

Green, A (2004) Investigation of range of pronunciation scale used in homogeneous and heterogeneous language groups, internal report 595, Cambridge ESOL.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C & Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing Vol 15, Cambridge: UCLES/ Cambridge University Press.

The CPE Textbook Washback Study

ROGER HAWKEY, CONSULTANT FOR CAMBRIDGE ESOL

Context and purpose

Increasing attention is being paid to test washback and impact studies as part of the test validation process (see, for example, Shohamy 2000, Hamp-Lyons 2001, Cheng and Watanabe 2004, Alderson 2004, Green 2005, Hawkey forthcoming). Cambridge ESOL's research and validation model has impact structured into it, of course. The model is, according to the Cambridge ESOL website (www.CambridgeESOL.org), 'designed to ensure that all ESOL assessment products meet acceptable standards in relation to the following four essential qualities:

- **Validity** – the extent to which test scores can be considered a true reflection of underlying ability
- **Reliability** – the extent to which test results are consistent and accurate, and therefore dependable
- **Impact** – the effect which a test has on candidates and other users, including society more broadly
- **Practicality** – the extent to which a test is practicable in terms of the resources needed'.

As candidature rises for language tests such as those in Cambridge ESOL's product range, so does the importance of the study of the washback of preparation courses for international English language tests, and of the textbooks designed for use on such courses (e.g. Saville and Hawkey 2004, Smith 2004).

In this article, washback is seen as part of impact, as in Hamp-Lyons (2000: 586). She acknowledges that Alderson and Wall's '*limitation of the term "washback" to influences on teaching, teachers, and learning (including curriculum and materials) seems now to be generally accepted, and the discussion of wider influences of tests is codified under the term "impact"* (Wall 1997), *which is the term used in the wider educational measurement literature'.*

The CPE Textbook Washback Study (Hawkey 2004) was commissioned by Cambridge ESOL. The purpose of the study was to test the hypothesis that *the constructs and content of the CPE test have washback effects on test preparation textbooks.* The study also sought answers to two research questions:

- To what extent has the CPE revision impacted on textbooks designed for use with CPE students?
- Are the changes in the exam reflected in the books?

Project design and instrumentation

Ten CPE-related textbooks were selected for the study. These included:

- four books designed for use in the preparation of candidates for the *pre*-2002 CPE exam
- four revised editions of these books designed for learners likely to take the revised CPE exam (see Weir and Milanovic 2003)
- two completely new books oriented towards the revised CPE exam.

Each of the selected books was rated independently by two language-teaching specialists selected, by Cambridge ESOL Main Suite staff, for their experience with teaching and/or testing activities related to the CPE exam and their other relevant expertise. A total of 20 textbook evaluations were made, distributed equally across the ten evaluators, in the pattern indicated in Figure 1.

Figure 1: Evaluator: text book rating schema

	Evaluators									
Books	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A previous version (prev.)		x	x							
***A** revised (rev.)*		x	x							
B (prev.)				x			x			
***B** (rev.)*				x			x			
C (prev.)					x					x
***C** (rev.)*					x					x
D (prev.)								x	x	
***D** (rev.)*								x	x	
E new	x					x				
F new	x					x				

The data collection instrument chosen for the evaluation of the textbooks was the Instrument for the Analysis of Textbook Materials (IATM). The original draft instrument was developed by Bonkowski (1996), then piloted and validated for use in the Cambridge ESOL IELTS Impact Study (Saville and Hawkey 2004, and Hawkey forthcoming). Smith (2004) also uses the IATM in an IELTS funded-research study of the accessibility of IELTS General Training Modules to 16–17 year old candidates.

The IATM was adapted for use in the CPE Textbook Washback Study on the basis of suggestions made by five members of the Cambridge ESOL Examinations and Assessment Group with a close knowledge of the constructs and content of the revised CPE exam. The IATM seeks both quantitative and qualitative information on:

- evaluator profiles and views of the CPE exam
- textbook characteristics, i.e.:
  - units of organisation
  - enabling skills
  - genres
  - language features
  - task types

The instrument then invites qualitative comment on a book's treatment of language skills and use of English, the quality of the book as a whole, and its relationship with the CPE exam.

Figure 2 exemplifies in comparative graph form the inter-rater reliability analyses carried out in the course of the study, in this case on the responses of two evaluators of the same textbook to IATM Item 4 on the *enabling skills* covered in the books (numbered 4.1 to 4.18 in the Figure). Where the two graph lines can be seen separately, there is disagreement on the coverage of the enabling skills concerned. Where only one line is visible, there is agreement. The pair of evaluations here, it will be seen from the graph lines, show quite strong agreement, as did most of the inter-rater reliability analyses.

Main findings

The hypothesis that the pre-revision and revised CPE exams exert strong washback on the evaluated textbooks in their treatment of language skills, micro-skills, task types, language elements and topics is supported by the Study. Further main conclusions are as follows:

- The evaluators note and consider it appropriate that the books tend to represent *directly* the content, approaches, activities and tasks of the exam.
- The evaluators consider that the textbooks concerned should also present opportunities and materials to develop learners' language knowledge and performance as appropriate to their individual levels, needs and interests.
- The evaluators consider that the revised and new editions of the preparation books reflect significantly the changes in the revised CPE exam.
- The revised and new CPE preparation books tended to be rated somewhat more highly than the pre-revision editions.

A longer, more detailed paper on the purpose, design, data management and analysis, findings and conclusions of the CPE Textbook Washback Study is being prepared for submission to a refereed journal.

References and further reading

Alderson, C (2004) Foreword to Cheng and Watanabe (Eds).

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

Bonkowski, F (1996) Instrument for the assessment of teaching materials, unpublished manuscript, Lancaster University.

Green, A (2005) Staying in Touch: tracking the career paths of CELTA graduates, *Research Notes* 19, 7–11.

Hamp-Lyons, L (2000) Social, professional and individual responsibility in language testing, *System* 28, 579–591.

Hawkey, R (2004). *Cambridge ESOL CPE Textbook Washback Study: Full report*, Cambridge: University of Cambridge ESOL Examinations.

—(forthcoming) *The theory and practice of impact studies: Messages from studies of the IELTS test and Progetto Lingue 2000.*

Liying Cheng, and Watanabe, Y (Eds) (2004) *Washback in Language Testing: Research Contexts and Methods.* New Jersey: Lawrence Erlbaum Associates.

Saville, N and Hawkey R (2004) The IELTS Impact Study: Investigating Washback on Teaching Materials, in Cheng and Watanabe (Eds).

Shohamy, E (2000) *The Power of Tests: a Critical Perspective on the Uses and Consequences of Language Tests*, Harlow: Longman.

Smith, J (2004) IELTS Impact: a study on the accessibility of IELTS GT Modules to 16–17 year old candidates, *Research Notes* 18, 6–8.

Wall, D (1997) Impact and washback in language testing, in Clapham, C (Ed.) *The Kluwer Encyclopaedia of Language in Education*, Vol 7, *Testing and Assessment*, Kluwer, Dordrecht, 334–343.

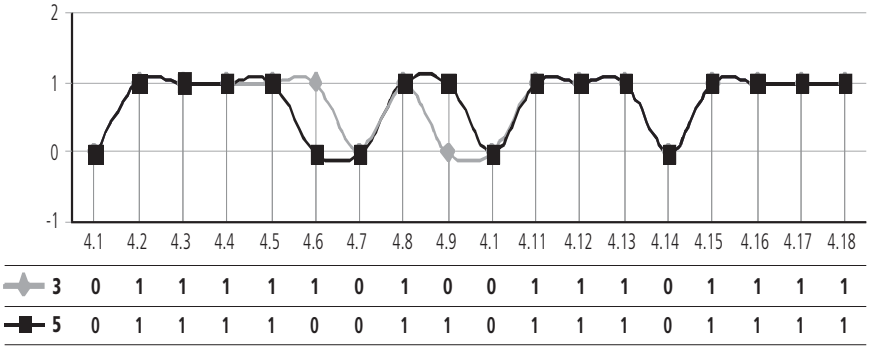
Weir, C and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Cambridge: UCLES/Cambridge University Press.

Figure 2: Comparisons between two evaluators (evaluators 3 and 5) on enabling skills covered in one book

Key:

1 = skills included

0 = skills not included



Testing language learners with special needs: sharing good practice

Over the years Cambridge ESOL has developed considerable experience, expertise and insight in how to make our examinations accessible to language learners with special needs. Cambridge ESOL was therefore pleased to be invited by the University of Jyvaskala, Finland, to contribute to a survey report on the teaching of languages to learners with special needs. The survey was being undertaken for the European Commission and had its roots in the European Year of People with Disabilities 2003. Its aim was to collate experience on good practice around Europe in the teaching and learning of languages among learners with special needs, at both policy and classroom level. In the absence of relevant Europe-wide quantitative data, the authors adopted a qualitative approach by interviewing and inviting contributions from a range of different stakeholders, including language test providers such as Cambridge ESOL and our partners in the Association of Language Testers in Europe (ALTE).

Cambridge ESOL's contribution to the report focused on Testing and is located in Chapter 2 of the report, *Insights*. Ruth Shuter, our Special Circumstances Co-ordinator, provided a comprehensive description of policy and practical issues when extending examination access to candidates with particular requirements. To do this, she drew on provisions and procedures used by several of the ALTE partners – Centre Internationale d'Etudes Pedagogiques (France), Instituto Cervantes (Spain) Goethe-Institut (Germany) and Cambridge ESOL (UK). The ALTE partners recognise that second/foreign language examinations can represent an incentive for learning and have a positive impact on teaching; they can also open doors to educational and employment opportunities. In the light of this, they strive to ensure that they have systems in place which allow special arrangements to be made for candidates with particular requirements due to temporary or permanent disability. In addition to the moral imperative, there is a growing legal obligation in many countries to protect people from discrimination on grounds of disability.

The purpose of the report for the European Commission is 'to examine the situation "on the ground" and make recommendations accordingly'. The authors have included a considerable amount of 'grassroots' level expertise because it reflects the extent to which localised solutions are being actively explored and implemented. They hope that the report will assist in the sharing of good practice and so provide a sound basis for future discussion and policy making in this area.

The final report was published in January 2005 with the title *Special Educational Needs in Europe – The Teaching and Learning of Languages – Insights and Innovation*. In February 2005 a formal presentation of the report was made to the 25 state representatives at the Commission which was apparently well received. The contributors are being encouraged to take up and further contribute to concrete development proposals, ideally on a trans-European basis. The report is publicly available via the Europa site at: http://www.europa.eu.int

IELTS Joint-funded Research Program Round 11: call for proposals

All IELTS-related research activities are co-ordinated as part of a coherent framework for research and validation. Activities are divided into areas which are the direct responsibility of Cambridge ESOL, and work which is funded and supported by IELTS Australia and the British Council.

As part of their ongoing commitment to IELTS-related validation and research, IELTS Australia and the British Council are once again making available funding for research projects in 2005/6. For several years now the partners have issued a joint call for research proposals that reflect current concerns and issues relating to the IELTS test in the international context. A full list of funded research studies conducted between 1995 and 2001 (Rounds 1–7) appeared in *Research Notes* 8 (May 2002) and studies conducted between 2002 and 2005 (rounds 8 to 10) are listed overleaf. Such research makes an important contribution to the monitoring and test development process for IELTS; it also helps IELTS stakeholders (e.g. English language professionals and teachers) to develop a greater understanding of the test.

All IELTS research is managed by a Joint Research Committee which agrees research priorities and oversees the tendering process. In determining the quality of the proposals and the research carried out, the Committee may call on a panel of external reviewers. The Committee also oversees the publication and/or presentation of research findings.

Details of the call for proposals including application forms and guidance on topics and resources can be found on the IELTS website: http://www.ielts.org

Studies funded under rounds 8–10 of the British Council/IELTS Australia joint funded research program

Round/Year	Topic	Researchers
Eight/2002	An investigation of the lexical dimension of the IELTS Speaking Test	John Read & Paul Nation, Victoria University of Wellington, New Zealand
	An examination of candidate discourse in the revised IELTS Speaking Test	Annie Brown, The University of Melbourne, Australia
	An empirical study on examiner deviation from the set interlocutor frames in the IELTS Speaking Test	Barry O'Sullivan & Lu Yang, University of Reading, UK
	Does the computer make a difference? Reactions of candidates to a CBT versus traditional hand-written form of IELTS Writing component: effects and impact	Cyril Weir & Barry O'Sullivan, University of Surrey, Roehampton, UK
Nine/2003	What makes a good IELTS speaking test? Perceptions of candidates and examiners	Christopher Hampton & Huang Chun, British Council Shanghai, China
	Student Identity, Learning and Progression: with specific reference to the affective and academic impact of IELTS on 'successful' IELTS students	Pauline Rea-Dickins, Richard Kiely & Guoxing Yu, University of Bristol, UK
	Exploring difficulty in speaking tasks: an intra-task perspective	Barry O'Sullivan, Cyril Weir & Tomoko Horai, University of Surrey, Roehampton, UK
	An investigation of the effectiveness and validity of planning time in part 2 of the oral module	Catherine Elder, University of Auckland, New Zealand, and Gillian Wigglesworth, University of Melbourne, Australia
	An examination of the rating process in the IELTS Speaking Test	Annie Brown, University of Melbourne, Australia
	A study of the linguistic and discoursal features in the output from IELTS Academic writing tasks	M A Yadugiri, consultant, formerly at Bangalore University, India
	Attitudes of tertiary key decision-makers towards English language tests: a New Zealand case study	Hilary Smith & Stephen Haslett, Systemetrics Research New Zealand and Massey University, New Zealand
Ten/2004	The use of IELTS for university selection in Australia: a case study	Kieran O'Loughlin, University of Melbourne, Australia
	Documenting features of written language production typical at different IELTS band levels	Florencia Franceschina & Jayanti Banerjee, University of Lancaster, UK
	The interactional organisation of the Speaking Test	Paul Seedhouse, University of Newcastle upon Tyne, UK and Maria Egbert, University of Southern Denmark, Sønderborg
	Exploring rater response-mode bias: students' writing on computer and pen-and-paper	Barry O'Sullivan, Roehampton University, UK
	An ethnographic study of classroom instruction in an IELTS preparation program	Peter Mickan, University of Adelaide, Australia
	The significance of socio-linguistic backgrounds of teachers of IELTS preparation courses in selected Malaysian institutions	Anne Swan, University of South Australia and Carol Gibson, consultant
	IELTS as a predictor of academic language performance	David Ingram, Amanda Bayliss & Andrea Paul, University of Melbourne, Australia

Recent publications of interest

One clear sign that language tests have 'impact' is the existence of various types of test-related publication, ranging from more theoretically-oriented books on assessment to very practical materials for test preparation. Ideally, the impact of such publications should be positive, i.e. they should support and encourage good quality language teaching/learning, and should help test-takers and other test stakeholders gain a sound understanding of assessment principles and the complex role testing plays in education and society. Details of three recent publications of interest are reported here, all of which have links – in one way or another – to Cambridge ESOL examinations.

Studies in Language Testing – Volume 16

Over the years, many different organisations in Britain have been involved in the testing and certification of English as a Foreign Language. For a variety of reasons some of these organisations no longer operate and, sadly, there is rarely any significant record of what they did or how they did it. Volume 16 in the Studies in Language Testing (SILT) series is entitled *A Modular Approach to Testing English Language Skills* and it was written in order to capture the history of the Oxford-ARELS English examinations and those of the Royal Society of Arts (RSA). The Oxford-ARELS and the RSA English examinations made an important contribution to the testing of English as a Foreign Language in the UK and around the world in the second half of the twentieth century. The volume also describes how these examinations impacted on the development of a new Cambridge ESOL examination – Certificates in English Language Skills (CELS).

From the 1980s onwards, the number of examination boards operating in the context of school examinations in the UK decreased, mainly for reasons related to government policy and the economics of running examination boards. The University of Cambridge Local Examinations Syndicate (UCLES) remains as the only university directly involved in school examinations in the UK. Through the 1980s and into the 1990s, UCLES formed strategic alliances with a number of other boards culminating in 1998 with the formation of Oxford, Cambridge, RSA (OCR). The drive to form OCR was a consequence of government legislation in the UK. As these various mergers took place, EFL examinations were sometimes involved but EFL was never the prime driver for any of the mergers. The consequence, however, was that by the late 1990s Cambridge ESOL (or UCLES EFL as it was then known) was in possession of a fairly large number of English language examinations that did not sit easily together and required significant resources to support them. A decision was made to produce a new suite, CELS, which aimed to rationalise the number of examinations offered while attempting to bring together the best features, as far as it could, of the examinations that were to be phased out.

In *A Modular Approach to Testing English Language Skills* Roger Hawkey begins by describing the English language teaching and testing context out of which the Oxford-ARELS and RSA examinations grew. He outlines succinctly a number of trends and evolves a very useful framework for the evaluation of communicative tests that he later applies to his analysis of the various examinations described in the book. He traces in some detail the history of the Oxford-ARELS and RSA examinations respectively. Although the records were sometimes sparse, Hawkey was able to gain access to a certain amount of useful documentation. However, what makes this volume so special are the numerous interviews that the author was able to conduct with many of the key people involved in the initial development and subsequent production of these examinations. He draws a fascinating, accurate and sympathetic picture of how the boards operated and how the examinations were conceived and subsequently produced. Hawkey's analysis helps us appreciate the great dedication and commitment of the individuals involved in their development and extensive appendices allow readers to get a very clear idea of what these examinations looked like. Many readers will find this of significant interest.

The volume brings us up to the present day by describing in detail the rationale and development of CELS. There is a significant focus on the validation of the new examination specifically on the validity, reliability, impact and practicality issues that surround examination development. The Cambridge ESOL approach to examination development gets significant attention and provides the reader with a very detailed understanding of the process and issues involved in question paper production and the management of change. The new CELS examination is compared to the Certificates in Communicative Skills in English (CCSE), one of the examinations that CELS replaced.

Roger Hawkey has produced a well written and fascinating history of a number of examinations that no longer exist, as well as a detailed review of CELS, the new examination that replaces them. He brings out clearly the high degree of professionalism that has characterised the British approach to English language testing over the years and illustrates well the quality of the new CELS and the great emphasis that Cambridge ESOL places on all aspects of the examination revision, development, production and validation process.

Volume 16 is the second historical survey in the SILT series, the first being Volume 15, which documented the revision of the Certificate of Proficiency in English (CPE). Volumes on the development of business English, academic English and English for young learners are currently in preparation. For more information on titles in the series go to:
www.cambridgeesol.org/research/silt.htm

Language Testing and Validation: an evidence-based approach

Professor Cyril Weir is well known for his considerable contribution to the field of language testing through books such as

Communicative Language Testing (1990), *Understanding and Developing Language Tests* (1993), and *Reading in a Second Language* (1998). His latest book – *Language Testing and Validation: an evidence-based approach* – aims to offer teachers and researchers a useful framework to enable them to evaluate critically the tests they encounter, whether these are tests devised for the classroom context or those provided by large examination boards, such as Cambridge ESOL.

Part 1 of the book maps out the types of validation evidence needed to give confidence that the results of performance on a test provide an accurate picture of the underlying abilities or constructs that are being measured. Part 2 provides real examples and procedures taken from tests around the world, including some of Cambridge ESOL's tests, and provides an evidence-based validity framework for asking questions of any exam or form of assessment. Part 3 suggests a number of research activities, large and small-scale, for generating data on whether a test matches up to the various criteria in the framework. This section will be particularly useful for Masters/professional teaching students undertaking research as part of their studies, as well as for practising teachers keen to put the framework into action. Part 4 backs up the discussion of research and practice with information on key electronic and paper-based resources.

As Weir acknowledges in his volume, Cambridge ESOL funded part of the research upon which the book is based and the socio-cognitive framework he proposes is currently providing us with a valuable heuristic for considering the various dimensions of validity as they relate to the Cambridge ESOL examinations.

Language Testing and Validation: an evidence-based approach is published by Palgrave Macmillan in their Research and Practice in Applied Linguistics series (www.palgrave.com).

Common mistakes …. and how to avoid them

Cambridge University Press has recently published a series of four booklets designed for students and teachers preparing for the PET, FCE, CAE and CPE examinations. The booklets are entitled *Common mistakes at PET/FCE/CAE/CPE* and they highlight common mistakes made by learners at these proficiency levels, offering guidance and practice on how to avoid them. They include exam-style exercises to familiarise students with the test format as well as regular tests to help them monitor their progress.

The *Common mistakes…* series is based on detailed analyses of the content of the Cambridge Learner Corpus (CLC). The CLC is a unique collection of over 55,000 anonymised examination scripts written by students taking Cambridge ESOL English exams around the world. It has been developed jointly by Cambridge ESOL and Cambridge University Press since 1993 and currently contains around 20 million words from different Cambridge examinations across a range of proficiency levels. Each writing script is coded with information about the learner's first language, nationality, level of English, age, etc. A unique feature of the CLC is that over 8 million words have been coded with a Learner Error Coding system devised by Cambridge University Press. This makes it possible to see words or structures that produce the most errors in Learner English at particular proficiency levels and it is analysis of the error-coded CLC content which has informed the development of the *Common mistakes…* booklets.

As the CLC grows in size and scope, it is proving an increasingly valuable tool for authors, editors and lexicographers working with Cambridge University Press who use it to develop dictionaries and ELT course books. Cambridge ESOL use analyses of the CLC content to inform a wide range of its test development, validation and other research activities, including: wordlist development and revision; monitoring of standards over time or across proficiency levels; and informing the development of materials and task formats at suitable levels of difficulty. The CLC will continue to provide a range of opportunities for future research, for example in the area of investigating the nature of lexis across the proficiency continuum.

For more information, go to:
http://www.cambridge.org/elt/commonmistakes

IELTS Masters Award 2005

Each year the IELTS partners – University of Cambridge ESOL Examinations, British Council, and IELTS Australia – sponsor an annual award of £1000 for the Masters Level dissertation or thesis which makes the most significant contribution to the field of language testing.

Details of the application process for the IELTS Masters Award 2005 can be found on the IELTS website – www.ielts.org
Please note that submission details may change from year to year and it is therefore important that the most current procedures are consulted.