Research Notes

Contents

Editorial Notes

Welcome to issue 27 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL. This first issue of 2007 brings a 'new look' for *Research Notes*. We hope you will find the refreshed design attractive to look at and easy to read, as well as a more convenient reference resource now that the contents list appears on the back as well as the inside front cover. We would welcome your comments on the new approach.

In this issue we focus on the theme of testing English for business and other work-related contexts. In his opening article, David Thighe discusses Cambridge ESOL's response to the changing assessment requirements that are resulting from globalisation and migration. He describes the growing demand for English language tests that are tailored to the needs of populations in various work-oriented contexts, outlining some of the principles that underpin the domain-related tests we offer, such as BEC, BULATS, ILEC and ICFE. Key issues include the notion of specificity, the nature of authenticity and the role of content knowledge. Questions of specificity and domain-related content also surface in the article by Jason Street and Kate Ingham, who describe the process of compiling, validating, and publishing word lists for our BEC Preliminary, PET and KET examinations.

Just as computer and other technologies are revolutionising the workplace and international business communications, so they are impacting increasingly on the testing and assessment of English within these contexts. The role of technology in developing a computer based test such as CB BULATS is explored by Louise Maycock, who describes the development of a tool that allows us to examine how an adaptive computer test functions. On a related theme, Neil Jones and Louise Maycock address the issue of comparability between the computer based mode and paper based mode of tests such as BEC and BULATS. Stuart Shaw describes efforts to conceptualise Cambridge ESOL's Writing assessment as a workflow in terms of different facets within an Electronic Script Management (ESM) environment; he shows how this developing technology draws on databases and can benefit the assessment of writing performance.

Steve Murray reports on Cambridge ESOL's recent initiatives to ensure the cultural accessibility of our examination materials in a constantly changing international context. Issue 27 includes a brief update on recent developments in the Asset Languages project as well as a conference report on Cambridge ESOL's involvement in the 2006 UK Language Testing Forum, which took English for Specific Purposes as its theme. Finally we report news of the winner of the 2006 IELTS Masters Award.

Editorial team for Issue 27: Lynda Taylor, Louise Maycock, Fiona Barker and Rowena Akinyemi.

Cambridge ESOL and tests of English for Specific Purposes

DAVID THIGHE RESEARCH AND VALIDATION GROUP

Introduction

In his recent work on the future of English, Graddol (2006) reports that there is likely to be an increase in the demand for English language skills from such groups as students, migrant workers, and from populations in nations that are emerging as future economic powers. What all these groups share is that the context in which they wish to use their English is likely to be work oriented. It is therefore conceivable that in the next few years there will be an increasing demand for work-oriented tests of English, i.e. tests containing language and tasks that are found in either quite specific work environments such as the law or call centre employment, or more general business English contexts. The 2006 Language Testing Forum held in Reading last November took as its main theme *Testing Language for Specific Purposes: the Future?* in recognition of the strong focus currently on this aspect of language assessment.

Cambridge ESOL's Business English tests

Since the early 1990s Cambridge ESOL have developed a number of English language proficiency tests with business students and employees in mind.

The Business English Certificate (BEC) suite of tests covers the Common European Framework of Reference (CEFR) levels B1, B2 and C1, providing comprehensive assessment of English within the business environment with components in Reading, Writing, Listening and Speaking. O'Sullivan (2006) details the development of and subsequent revisions to the BEC suite and also examines other work-related language tests. O'Sullivan raises some important issues with regard to Language for Specific Purposes (LSP) tests generally and several of his defining features of LSP tests are explored briefly in this article.

The Business Language Testing Service (BULATS) is a comparatively shorter test than BEC, aimed at businesses who need to locate quickly and reliably their employees' ability to communicate in English across a range of CEFR levels. Four different test modules are currently available: the BULATS Computer Test for Reading and Listening (computer-based); the BULATS Standard Test for Reading and Listening (paper-based); the BULATS Writing Test; and the BULATS Speaking Test.

The International Legal English Certificate (ILEC) and the International Certificate in Financial English (ICFE) are aimed at students and employees in the legal and financial work environments respectively. These are something of an innovation for Cambridge ESOL in that they test language used in specific areas of the business domain and as a result raise their own particular concerns. Articles in previous issues of *Research Notes* described the approach to developing these tests (Thighe 2006, Ingham and Thighe 2006) and the role of content or subject experts in their original design and on-going production (Corkill and Robinson 2006).

Issues in ESP testing

Developing English for Specific Purposes (ESP) tests such as those described above raises issues that are distinct from those involved in more general purpose English tests; the remainder of this article summarises these issues, offering a brief discussion about each one.

In his work on the testing of Language for Specific Purposes (LSP), Douglas (2000) describes an LSP test as

> ...one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain. (Douglas 2000:19)

Douglas argues that authenticity is central to LSP testing. The underlying assumptions here are that

1. language use and performance can vary according to context,
2. specific contexts have distinguishing lexical, semantic, syntactic and phonological features, and
3. these features can only be realistically approached through the use of authentic test material.

O'Sullivan (2006:3) provides a brief but useful review of the research that substantiates the first two of these claims.

Content/background knowledge

Once authenticity is placed at the centre of the ESP assessment enterprise, various issues arise for those involved in the practical production of ESP tests. Chief amongst these is the role of content or background knowledge in the test. Content knowledge is the largely non-linguistic knowledge that a candidate brings to a test; for example, for a candidate to answer a task about restaurants it is necessary for that candidate to know what restaurants are, how they work etc. With general English tests much time is spent ensuring that content knowledge is not tested; this is usually done by providing themes and contexts that all candidates, regardless of cultural background, can reasonably be expected to be familiar with.

For ESP tests the situation is more complex, for at least two reasons. ESP tests targeted on one work domain must

include contexts that are appropriate for all people working in that domain. For example, in the development of the International Legal English Certificate it was necessary to ensure that lawyers were not at a disadvantage if they were only trained in the Anglo-American common law system and not civil law. As this content element of test tasks is either inaccessible to assessment specialists or very difficult for them to evaluate without assistance from another group of specialists, it proved necessary to develop an iterative, ongoing relationship with content experts to ensure the fairness of the test. This is explained in more detail in Corkill and Robinson (2006).

The issue of content knowledge across different sub-domains can be thought of in terms of a 'horizontal' dimension; a second issue constitutes a 'vertical' dimension, i.e. the degree of content knowledge amongst candidates. In relation to this aspect, we cannot, as we do in tests of general English, assume that all candidates have an equal level of content knowledge to deal with the task adequately. Indeed, the idea of an 'adequate' or cut-off level appears misplaced in an ESP environment. This perhaps becomes clearer when we consider our own working environment and output, where we are continually performing tasks poorly or well according to our level of expertise in a particular area of our work. Clapham's (1996) research into the effect of candidate content knowledge on a test of English for Academic Purposes showed that content knowledge can have an impact on scores. For intermediate candidates there was a correlation between content knowledge and test score. This is what Douglas (2001) succinctly labels the issue of *inseparability*: we must expect in ESP tests that content knowledge will play a part in the candidate's performance; if it does not, then it is questionable whether the test is really an ESP test. As O'Sullivan writes:

> It can be argued that a test of language for a specific purpose should not even try to avoid the background knowledge issue, as it is this that defines the test. How we deal with the situation will depend on the degree of specificity of the test and the inferences we intend to draw from performance on the test. (O'Sullivan 2006:5)

The notion of specificity

Clapham (1996) also noted, and it is referred to in the quote above, that the role of candidates' content knowledge was related to the specificity of the test; in other words, the more specific an ESP test is, the more content knowledge and language knowledge are interconnected. This concept of a cline of levels of specificity is taken up by O'Sullivan and is illustrated in Figure 1. Here a completely specific language test is defined as one that would focus only on language unique to a specific use domain. As the specificity of the test increases so does the degree to which content knowledge is tested. Also, as specificity increases, the generalisability of the test results decreases. We may posit that examinations such as the First Certificate in English are non-specific, while examinations with a business orientation such as BEC are more so. An examination of English in a legal environment such as ILEC is likely, if it is well constructed, to be even more specific.

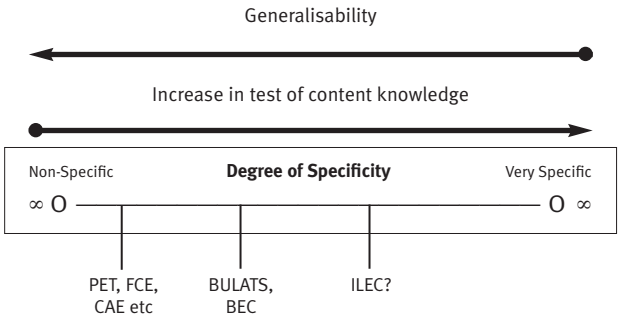

Figure 1: The specificity continuum in ESP tests

Situational and interactional authenticity

O'Sullivan's contribution to the debate on the nature of ESP testing is to remind us of the importance of the distinction between *situational* and *interactional* authenticity. Situational authenticity is achieved when test tasks resemble tasks in the target language use situation. Interactional authenticity is achieved when the mental processing of the candidate in the test resembles that of the candidate in the target language use situation (see also Bachman and Palmer 1996). This view chimes well with elements of a systematic approach to test validation which has been developing recently, referred to as the Socio-Cognitive Framework and presented in Weir (2005). Two particular components of the framework – *context-based validity* and *cognitive validity* – reflect in some sense the earlier notions of situational and interactional authenticity.

For O'Sullivan interactional authenticity should be the goal of all Language for Specific Purposes (LSP) tests, including ESP tests, but the question is how to measure this.

> If we show that the cognitive processes involved in an LSP test task performance reflect those of the specific language use domain they are designed to reflect, then we can claim with some confidence that our test task demonstrates interactional authenticity. It is quite possible that such processing may well differ in important respects from general purpose task performance… By demonstrating differences in internal processing between LSP and general purpose tasks we are offering an additional argument in favour of the distinguishability of language use domains. (O'Sullivan 2006:183)

O'Sullivan sees the key for assessing situational and possibly interactional authenticity in defining test specificity as multi-dimensional and, importantly, measurable. He provides a hands-on method for examining test specificity that Cambridge ESOL is currently exploring in relation to our own ESP tests. This will allow us to ascertain whether certain components or tasks within a test have the required level of specificity from a number of different facets such as, for example, lexical precision, structural range and text length.

Conclusion

This outline of current thinking on ESP testing has shown that developing such tests raises a number of critical issues, such as the role of content experts in test production, levels of test specificity, and the inseparability of content and language knowledge; although

methodologies for addressing these issues are steadily emerging, they continue to present language testers with theoretical and practical challenges. Other issues such as the role of content specialists in devising marking criteria (Douglas's indigenous assessment criteria, Douglas 2000) and placing ESP tests on a largely non-context based framework such as the CEFR remain to be dealt with. As the demand grows for tests of ESP, there will be plenty for language testers to engage with.

References and further reading

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

Clapham, C M (1996) *The Development of IELTS: A study of the effect of background knowledge on reading comprehension* (Studies in Language Testing, volume 4), Cambridge: UCLES/Cambridge University Press.

Corkill, D and Robinson, M (2006) Using the global community in the development of ILEC, *Research Notes* 25, 10–11.

Douglas, D (2000) *Assessing Languages for Specific Purposes*, Cambridge: Cambridge University Press.

Douglas, D (2001) Three problems in testing language for specific purposes: authenticity, specificity, and inseparability. In Elder, C, Brown, A, Grove, E, Hill, K, Iwashita, N, Lumley, T, McNamara, T, and O'Loughlin, K (Eds), *Experimenting with Uncertainty: Essays in honour of Alan Davies* (Studies in Language Testing, volume 11), 45–52, Cambridge: UCLES/Cambridge University Press.

Graddol, D (2006) *English Next: Why Global English may mean the end of 'English as a Foreign Language'*, The British Council, Available at: www.britishcouncil.org/learning-research-englishnext.htm

Ingham, K and Thighe, D (2006) Issues with developing a test in LSP: the International Certificate in Financial English, *Research Notes* 25, 5–9.

O'Sullivan, B (2006) *Issues in Testing Business English: The revision of the Cambridge Business English Certificate* (Studies in Language Testing) volume 17, Cambridge: Cambridge ESOL/Cambridge University Press.

Thighe, D (2006) Placing the International Legal English Certificate on the CEFR, *Research Notes* 24, 5–7.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Hampshire: Palgrave Macmillan.

Publishing vocabulary lists for BEC Preliminary, PET and KET examinations

JASON STREET AND KATE INGHAM ASSESSMENT AND OPERATIONS GROUP

Introduction

In March last year, Cambridge ESOL published vocabulary lists for the Business English Certificate (BEC) Preliminary, Key English Test (KET) and Preliminary English Test (PET) examinations on the ESOL website. The publication of these lists is a significant step. It is the first time a guide to the vocabulary content of an ESOL exam intended for adult learners has been produced since the publication in 1980 of the Cambridge English Lexicon (Hindmarsh 1980). The decision to place the vocabulary lists in the public domain was taken against a background of continuing efforts by Cambridge ESOL to provide stakeholders with more information about its examinations and at a time when testing organisations internationally are attempting to define more explicitly what language learners at particular levels can do.

Publication of the vocabulary lists is a consequence of the increased knowledge acquired in recent years by Cambridge ESOL about the vocabulary use of its candidates, as the result of the development, with Cambridge University Press (CUP), of the *Cambridge Learner Corpus* (CLC) (see Barker 2006). The CLC consists of electronically stored exam scripts of over 75,000 candidates from twenty different Cambridge ESOL Writing components, representing more than 23 million words, which can be accessed and analysed. The corpus operates with sophisticated Windows-based software which allows users to carry out a wide range of searches and concordances.

Background to the use of word lists

Word lists have been used in teaching and assessment for many years. Lists such as *A General Service List of English Words* (West 1953) and the *Cambridge English Lexicon* (Hindmarsh 1980) were developed more intuitively than empirically using the judgement of specialists: largely by counting the occurrence of words in texts selected as being representative. The Council of Europe's *Threshold* and *Waystage* vocabulary specifications (Van Ek and Trim 1991a and 1991b) were developed in relation to language functions. Recently, the creation of native speaker and learner corpora has made it possible to develop more empirically-based word lists. Expert judgement, however, was the principal guide to the original lists. Hindmarsh's claim that, in making decisions on word selection for his English Lexicon (1980: vi), the primary aides were 'intuition and wide experience of teaching' still carries some weight today. More recently, McCarthy and Carter have observed that 'corpus statistics can take us a considerable way from what intuition and conventional practice alone can provide, but the one should not exist without the other' (2003:7).

At Cambridge ESOL, vocabulary lists for the Lower Main Suite (LMS) General English examinations (KET and PET) and the BEC Preliminary examination have, in fact, been in existence for some time, but until now have been intended for the use of writers for the relevant question papers to inform the lexical content of those examination papers. The PET and KET lists were originally based on the *Threshold*

and *Waystage* lists, as was the BEC Preliminary list, with the inclusion of business vocabulary considered appropriate to the level. The business domain lexis was added following work carried out by external specialists in the field of Business English. The three lists have been gradually extended over the years through an annual review process, with the addition of items felt to be of need to candidates as a result of changes in usage in the language, and are directly informed by reference to the CLC (see description of the process below). Usage in vocabulary can change quickly, as is shown by the growth in the last decade of the use of technology-related language and associated lexis.

The annual review procedure

New words proposed for inclusion in the lists are subject to an established procedure used in the annual review of the content of all the word lists. (For a full description of the annual word list review procedure, see Ball 2002.) Firstly, the frequency of occurrence of these words is established in a range of native speaker corpora. For BEC Preliminary, the following corpora are used:

- British National Corpus (BNC) (written and spoken native speaker English including business oriented words)
- Business Texts Corpus (based on US and UK business articles)
- Trial Web Corpus (words of contemporary Business English taken from the internet).

Next, the list of suggested words is compared against CLC-derived lists that illustrate the written production of a large number of candidates taking BEC at different levels. Each word is tested against various frequency criteria and the resulting data guide the final discussion as to whether words are added to the list or not. The significance of the data is that if the CLC shows that candidates at BEC Vantage level are using a word productively in writing, this strengthens the case for its inclusion on the BEC Preliminary vocabulary list.

Nonetheless, data from learner corpora, if used as the *principal* basis for a word list, does have some limitations. Firstly, the CLC consists of the written output of exam candidates and such productive vocabulary does not provide a reliable guide to receptive vocabulary knowledge, which is likely to be different in size and nature. Secondly, the frequent occurrence of some words in the corpus may well be a result of 'task effect' – if candidates are asked to describe their family and home in a PET Writing task or make meeting arrangements in a BEC Preliminary Writing task then, naturally, we would expect to find a high occurrence of vocabulary related to these topics in the data gathered from such tasks. However, the CLC is only *one* of the corpora used to inform the vocabulary included in the lists. Furthermore, as the CLC is growing at a very fast rate, with data from recent candidate scripts added on a regular basis, any 'task effect' decreases with time.

Until early 2006 the lists were for reference use of question paper materials writers only. With the decision to make the lists publicly available, there was a concern to ensure regularity of format across all three. A Cambridge ESOL working group was therefore set up in 2005 with the aim of reviewing the BEC Preliminary, PET and KET item writer vocabulary lists prior to their publication on the ESOL Teaching Resources website.

Initial aims and considerations

A number of issues were discussed within the working group. Members of the group were in agreement on the need for simplicity: the word lists would need to be accessible to teachers from a wide variety of backgrounds. At the same time, the lists should be optimally helpful and display consistency of convention from list to list. To ensure that the lists were simple to use, lexical items were chosen as headwords. Only base forms would be listed and not their inflections. Common prefixes and suffixes would be listed in an appendix and words formed with them not included in the body of the main list itself. No examples of word use would be given, except where these served to distinguish words with multiple meanings.

There was lengthy discussion on how compounds and multi-verb or phrasal verbs could be most helpfully listed. It was finally decided that where the meaning of the compound was literal and formed from two words which already appeared on the list (e.g. *sports centre* and *bank account*), they would not be listed separately. Compounds where the meaning was not transparent (e.g. *shortlist*) or where it might be difficult to grasp because of cultural trappings (e.g. *left luggage*) were listed separately. Any multi-part verbs which were felt to be 'sum-of-parts', i.e. where their meaning could be clearly understood from their constituent parts (e.g. *pay back*) and where these constituent words were already included in the lists, were not included in the main list. Non-literal phrasal verbs (e.g. *put off* meaning *postpone*) were listed in the main body of the list. It was also agreed that the name of products which were, in effect, trademarks (e.g. *Coca-Cola*) should be removed.

A number of closed vocabulary sets, such as numbers, days of the week and names of countries would be included in appendices to all lists, rather than appearing within the main lists. There was also a decision to be taken about topic lists (such as *The Weather*) which existed as appendices to the PET and KET vocabulary lists. It was agreed that topic lists should be retained as appendices to the LMS lists as they may be found to be helpful to teachers. The BEC Preliminary list, however, did not include any topic lists as it was based mainly on the vocabulary of business, which was considered to be a specific domain in itself (Horner and Strutt 2004).

Each list would be prefaced by an introduction highlighting a number of points. Users would need to know that the lists were intended as reference tools, as guides to the vocabulary a candidate might need to know and recognise, and were not intended to be a definitive collection of all the lexical items an elementary or pre-intermediate student of English should know or encounter. A further caveat was that the vocabulary included covered that needed for both receptive and productive tasks on the particular examination. There was also a risk that some users might view the word lists as forming a vocabulary

syllabus[1]: some teachers might, for example, teach their students only the words the lists contained. Users were to be reminded that language learners need to be equipped with approaches to deal with unfamiliar vocabulary, and were referred to the teaching resources section on the ESOL website which contains advice on strategies learners should use in order to understand and learn new vocabulary when reading texts. Learners also need to develop their own personal vocabularies: to learn specific words and phrases which might not be in the word lists in order to talk about their individual lives and particular personal circumstances.

Methodology

The initial work on reviewing the word lists was carried out by external consultants from the BEC Preliminary, PET and KET item writer teams: people with the greatest familiarity with the lists and their contents. The content of the existing item writer vocabulary lists was checked carefully for inconsistencies: individual items were cross referenced against decisions on the format of the main list that had been reached, and if necessary, removed. It was discovered, for example, that a number of words included within one of the lists was covered by the appendix in which affixes appeared.

Finally, a comparison study involving the draft word lists was done using WordSmith Tools to check the three lists for overlap. The project aimed to investigate the degree of overlap between each list. Overlap was anticipated between all three lists, but given the different development of all three lists over time, the degree of overlap was unknown.

Findings

The comparison study in fact revealed a significant overlap between the KET, PET and BEC Preliminary lists. As shown in Table 1, 1061 words were common to the KET and PET lists and 1949 were common to PET and BEC Preliminary lists. This was not unexpected, as the lists did have common origins.

It did, however, leave a substantial minority of the words on the BEC Preliminary list without any overlap on the PET list. (55 words occurred only on the KET list, 754 words occurred only on the PET list, and 1004 occurred only on the BEC Preliminary list.) This may appear surprising as both the latter examinations represent B1 level on the Common European Framework of References (CEFR) for Languages.

This can partly be explained by the factor of domain knowledge. Business English (BE) covers a range of situations, and the domain knowledge which is assumed in Business English varies. Lexis is an important difference between the two and is partially related to this question. Business English includes lexis not included in General English (GE), firstly, words and phrases which refer to general concepts but which would not normally be used in everyday non-business situations, e.g. *purchase*, *objective*, *consumer*. These words can be considered as General English vocabulary at a higher level, for example, *purchase* is a BEC Preliminary word for Business English, but intermediate (FCE) for General English. Secondly, there are words for general business concepts which arguably fall outside the domain of General English at lower/intermediate levels, e.g. *invoice*, *export*, *freight*. Furthermore, the lexis found only on the BEC Preliminary list appeared to fall within a fairly limited number of semantic categories. These included institutions and places of business, money, business people and events, modes of communication, products and companies and lexis connected with technology. In addition, the overlap exercise revealed that the BEC list included very little lexis used to describe feelings, society, family, home, personal issues and personal activities but was concerned to a far greater extent with the 'impersonal'. The 1004 words which occur in the BEC Preliminary list only could be an indicator of business-focused B1 level vocabulary, but further research is needed to substantiate this.

Conversely, the PET list included lexis not covered by BEC Preliminary, in that there is a much more extensive range of vocabulary for clothes, food, animals, sport, health and entertainment. The following are examples of words appearing on the PET vocabulary list but not BEC Preliminary: *carrot*, *camel*, *volleyball* and *opera*.

There were also cases of the same word appearing on both lists but which needed to be exemplified differently to indicate that only a certain use of it would be expected in the respective examination. *Market* on the BEC list features as noun and verb parts of speech with examples as *There's a good market for fax machines here* and *They marketed their goods overseas*. On the PET list, *market* appears as noun only with the example of buying something from a market.

In March 2006, the three word lists were posted on the relevant exam page of the Cambridge ESOL Teaching Resources website. The BEC Preliminary list consisted of 2995 headwords; PET had 2708 and KET had 1111 (plus topic lists). The published lists include the vast majority of lexical items expected in these exams and, albeit to a more

1. The wordlists for the Young Learners' English (YLE) Tests do form a syllabus, but teachers are recommended nonetheless to teach words beyond the lists. See forthcoming *Research Notes* 28 for an article on the YLE wordlist review.

Table 1: Words occurring in more than one list

Lists compared	LMS		LMS vs BECP				All 3 lists		
	KET	PET	PET	BECP	KET	BECP	KET	PET	BECP
Matching items	1061		1949		834		809		
Total words per list	1111	2708	2708	2995	1111	2995	1111	2708	2995
% of each list	95.6	39.2	69.5	65.1	75.1	27.8	72.7	29.8	27.0

limited extent, the likely output of candidates. The selection of words that they contain is the result of the collective experience of ESOL external consultants and writers, feedback from examiners and course providers over a number of years and the analysis of data from native speaker corpora and the learner corpus, the CLC. As such, the word lists may be considered useful reference tools in the preparation of BEC Preliminary, PET and KET candidates, and in the development of course materials for these examinations.

Future developments

The content of the lists will continue to be reviewed annually and items which are possibilities for inclusion or removal will, as the CLC is developed further, be subject to an evermore sophisticated corpus-based analysis. More generally, the work on the word lists contributes to on-going Cambridge ESOL research into the differences between 'receptive' and 'productive' vocabulary, across different levels and for different types of English. Using a number of different corpora, including the CLC, and with reference also to ESOL's computerised bank of items for receptive data, LIBS (see Marshall 2006), it is hoped that it may be possible in the future to identify more clearly the range of words that an average candidate at any level should be able to recognise or produce (see Ball, 2002). This work will, in turn, feed into a larger collaborative project (English Profile) to produce a comprehensive reference level description for English.

A possible area to explore, one which emerged as a result of the comparison of the final word lists, is whether words occurring in one of the lists and not in another could help to better define 'Business' English at the B1 level. This could inform the on-going validation of the construct underpinning Cambridge ESOL Business English tests and examinations.

References and further reading

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

Barker, F (2006) Corpora and language assessment: trends and prospects, *Research Notes* 26, 2–4.

Cambridge ESOL Teaching Resources website: www.cambridgeesol.org/teach

Hindmarsh, R (1980) *Cambridge English Lexicon*, Cambridge: Cambridge University Press.

Horner, D and Strutt, P (2004) Analysing domain-specific lexical categories: evidence from the BEC written corpus, *Research Notes* 15, 6–8.

Marshall, H (2006) The Cambridge ESOL Item Banking System, *Research Notes* 23, 3–5.

McCarthy, M and Carter, R (2003) What constitutes a basic spoken vocabulary? *Research Notes* 13, 5–7.

Van Ek, JA & Trim, J L M (1991a) *Threshold 1990*, Strasbourg: Council of Europe Publishing.

Van Ek, J A & Trim, J L M (1991b) *Waystage 1990*, Strasbourg: Council of Europe Publishing.

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

WordSmith Tools website: www.lexically.net/wordsmith

Using simulation to inform item bank construction for the BULATS computer adaptive test

LOUISE MAYCOCK RESEARCH AND VALIDATION GROUP

Introduction

The BULATS (Business Language Testing Service) computer-based (CB) test assesses listening, reading and language knowledge skills in a business context. It is used primarily by organisations needing a reliable method of assessing the language proficiency of groups of employees or trainees in English, French, German or Spanish. CB BULATS is a Computer Adaptive Test (CAT), which means that it adapts to the level of the candidate based on their performance on the items administered so far. As Wainer (2000:10) points out, the idea of an adaptive test is to mimic what a 'wise examiner' would do – ask a harder question if the first is answered correctly or an easier one if it proves too difficult, based on the observation that we learn very little about the ability of an examinee if the test items are all too easy or too difficult. There are numerous advantages associated with using adaptive tests:

- Because items are targeted at the level of the candidate, more information is contributed by each item and the test is consequently more efficient. Hence, it is possible to attain the same level of precision with fewer items than are required on a linear test.
- Test takers are able to work at their own pace.
- Candidates are suitably challenged at an appropriate level, rather than being bored or frustrated by items which are too easy or too difficult for them.
- The test can be marked and scored immediately, so results are instantly available (though this advantage is not restricted to CATs, as this may be the case for many computer-based assessments).
- Security of materials is vastly improved because each candidate receives a different selection of tasks.

Computerised adaptive testing is an application of item banking. Cambridge ESOL has been using item banking for a number of years, and previous *Research Notes* articles have described our approach in detail (Beeston 2000, Marshall 2006). An item bank is a large collection of test items for which statistical features such as item difficulty

have been estimated (*calibrated*) and mapped onto a common scale. The techniques which make this possible derive from Item Response Theory (IRT), which Wainer (2000:13) refers to as 'the theoretical glue that holds a CAT together'. IRT comprises a family of mathematical models which are used to calculate the probability of a candidate of given ability answering a particular item correctly. The simplest of the IRT models is the Rasch model, which calculates this probability as a function of the difference between the ability of the candidate and the difficulty of the item. The use of Rasch analysis makes item banking possible by enabling all items to be calibrated on the same difficulty scale. Any combination of items can then be selected from the bank and, because the difficulty of each of those items is known with some precision, it is possible to estimate the ability of the candidate based on their responses to that selection of items. This is an essential requirement for CATs because, since each candidate receives a different selection of items, comparisons cannot be made on the basis of a 'number correct' type score. As well enabling ability to be estimated on completion of the test, IRT is also used throughout a CAT to estimate ability after each task has been administered, in order to inform selection of the next task.

The CB BULATS test algorithm

The CB BULATS test algorithm comprises three main components: task selection, stopping criteria, and ability calculation. As already explained, the ability of the candidate is estimated throughout the test, after the response to each task has been submitted. As more information is obtained the ability estimate is gradually refined, so that the Standard Error of Measurement (SEM) of the estimate is reduced throughout the duration of the test. Theoretically, a CAT would end once a desired level of precision has been reached, characterised by the SEM falling below a certain level. However, for practical reasons this criterion cannot be the only method of stopping the test as it would result in an indeterminately long test for some candidates who perform more erratically or have an ability level at either of the extremes of the distribution. The CB BULATS algorithm employs several criteria, so that the current section comes to an end when any one of the stopping rules have been satisfied. These criteria include a target SEM, maximum number of items and also a time limit which test administrators may choose to impose if they so wish.

Task selection

The selection of the next task to be administered is the most complex part of the CB BULATS algorithm. The most common methods involve selecting the task contributing the maximum amount of *information* at the current ability estimate of the candidate, or some similar statistical approach. For the Rasch model, the maximum amount of information is contributed by items with difficulty estimates closest to the ability estimate. However, whilst this strategy would be most efficient in terms of producing a test with maximum precision, there are other considerations relating to the validity and practicality of the test which must be dealt with by the algorithm. For example, balancing the range of task types administered is an important aspect of construct validity, and balancing the exposure of material is equally important for security of the item bank. In order to account for these considerations, the CB BULATS task selection procedure involves selecting a subset of potential tasks, which are of the right type and fall within a certain difficulty range, and multiplying together a series of weights which work to control material exposure, maximise targeting as far as possible and so on. A task is then selected randomly from the top few tasks, sorted in decreasing order of overall weight.

Construction of CAT item pools

The adaptive algorithm implemented in CB BULATS is clearly very complex but, as Flaugher (2000:38) points out 'the best and most sophisticated adaptive program cannot function if it is held in check by a limited pool of items, or items of poor quality'. This article is not intended to address the quality of items: CB BULATS material goes through the same rigorous quality control procedures as other Cambridge ESOL examinations, as detailed elsewhere (e.g. Marshall 2006). However, construction of item pools for CB BULATS is an important issue. The term *item pool* is used here to distinguish the material selected for a particular test version from the master *item bank*, which contains all of the material available for selection. New versions of CB BULATS are produced periodically, containing different test material and the potential to manipulate parameters of the adaptive algorithm if it is considered necessary.

Constructing an item pool for an adaptive test poses interesting problems not faced in the construction of linear tests where, given a sample size, the exposure of each item in the test is pre-determined and equal. In an adaptive test this is not the same because the exposure of a given task depends on a range of factors relating to the likelihood and frequency of its selection.

The primary considerations when constructing a new pool of material are that there is a sufficient range of task types across each of the levels for the test to perform well, and that the exposure of material is controlled by ensuring more material is available for the task types and levels which are likely to be utilised more often. Sometimes these considerations may be conflicting when constructing a pool of predetermined size. For example, Flaugher (2000:42) suggests that 'a satisfactory pool of items for adaptive testing is one characterized by items with … a rectangular distribution of difficulty'. Whilst this would ensure that sufficient material was available such that the test would perform equally well across all levels, this does not take into account constraints on exposure of material. Even if the test taker population had a rectangular distribution of ability, exposure would still be greater for the tasks in the middle of the difficulty range because this is where the tasks at the beginning of the test are selected from, until more information is obtained about the candidate's ability. If the population were normally distributed then exposure levels of mid-range tasks would be further increased

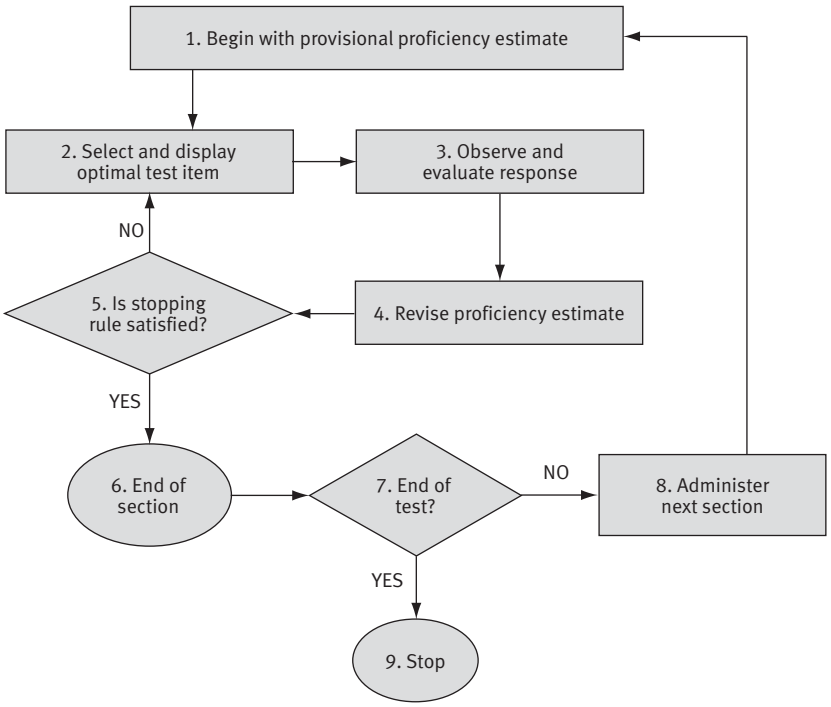

Figure 1: Adaptive test logic (adapted from Thissen and Mislevy, 2000:106)

because most of the candidates would fall in the middle of the range. Hence, to control exposure levels it would be preferable to construct the item pool such that most of the tasks fall in the middle of the difficulty range, with fewer tasks of extremely high or low difficulty. It is essential that these competing issues be considered when constructing a new item pool.

The CAT Simulator

When a new item pool has been constructed, checks need to be made to ensure that the test will perform sufficiently well for candidates of different levels and that exposure rates for any individual task will not be too high. Since the utilisation of tasks cannot easily be predicted, the only way to address these issues is by the use of simulation. The Cambridge ESOL CAT Simulation Tool has been designed to mimic precisely the algorithm implemented in CB BULATS. It works by generating a sample of candidate abilities, based on population characteristics specified by the user (i.e. the probability distribution and parameters of that distribution). The test is then simulated for each of the 'candidates' generated. Because their 'true' ability is known (since it was generated) the Rasch model can be used to calculate the probability of a correct response to a given item. In order to model the probabilistic nature of a candidate's response, a random number between 0 and 1 is generated. If the generated number is less than the probability of a correct response then the candidate is deemed to have answered the item correctly; otherwise an incorrect response is recorded. For example, say the probability of the candidate responding correctly was calculated to be 0.6. Next, generate a random number between 0 and 1. The probability of this number being less than 0.6 is 0.6, so this has the same probability of occurring as does a correct response: thus, record a correct response if the generated number is less than 0.6. This simulation approach is similar to that described by Luecht et al (1998).

Aside from the generation of candidate responses, everything else in the simulator – task selection, stopping rules, ability estimation and so on – is exactly the same as in the actual test. Simulating a large number of candidate performances therefore allows us to predict how the test will perform and how the item pool will be utilised.

Use of simulation information for item pool construction

After running a simulation a range of information can be extracted from the simulator in order for recommendations to be made regarding amendments to the item pool – for example, specific task types or levels where there is insufficient material or, indeed, a surplus of material (where it may be appropriate to retain some tasks for a future item pool).

Assessing the precision of the test

Targeting plots like that in Figure 2 are produced for candidates at each level, to show how the average difficulty corresponds to the average estimated ability during the test sequence. Error bars are plotted to represent the standard error of the ability estimates. Note that the test is in two sections: the Reading and Language Knowledge section ends after around 32 items and is then followed by the Listening section (hence, the noticeable dip in the middle of the plot as the current ability estimate returns to the centre of the distribution). Figure 2 shows the ability estimate steadily increasing for this group of candidates, and the item difficulties can be seen to be increasing correspondingly. Also, the error bars show how the SEM of the ability estimate is reduced during the course of each section.

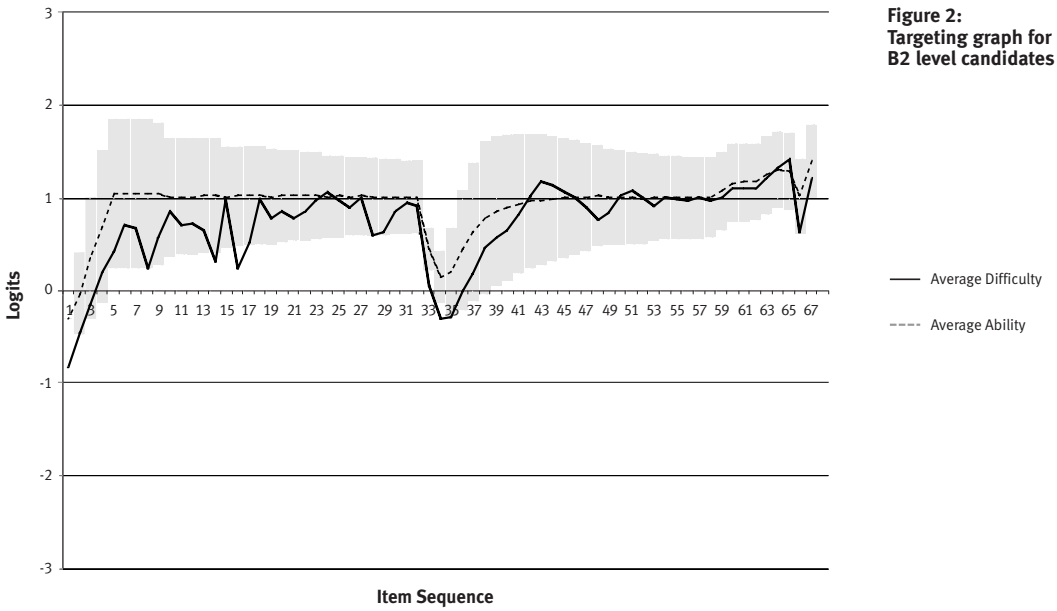

Figure 2: Targeting graph for B2 level candidates

The distribution of the SEM is compared for candidates at each level, and classification statistics, which indicate the percentage of candidates correctly classified into their 'true grade', are also produced. A plot of SEM over time (i.e. from the first candidate to the last) is also produced in order to ensure that constraints on material exposure, which effectively mean that there is more freedom for the system to select the better targeted items for earlier candidates, do not impact on the reliability of the test for later candidates. Together this information gives an overall picture of how well the test is performing for candidates at each of the levels and highlights if there is insufficient material at certain levels.

Algorithm performance

Features of algorithm performance are studied to ensure that the test will function as expected for a given item pool. This information includes summarising how the stopping criteria are used (e.g. what proportion of tests end because the desired level of precision is reached), the average number of items per section, how often a line of the test specification had to be skipped because there were insufficient tasks of the specified type available, and how far the selection of task types reflected the preference of a given task type in each part of the test specification.

Use of material

An important reason for simulation is to check the expected exposure levels for different tasks. Information on the minimum, maximum and average rate of exposure of tasks of each task type and level are produced. Types and levels where these figures are high represent areas where more material may be required in order to prevent over-exposure and reduce the risk of material security being compromised. Graphs of the distribution of task types for each level of candidate are also produced to ensure that all levels receive an adequate range of task types. Again, deficiencies in the item pool may be detected if the information produced is not as expected.

Conclusion

Simulation plays a vital role in the construction of item pools for adaptive tests, allowing careful assessments of likely test performance and material use to be made so that the contents of an item pool may be revised a number of times before being produced as a new CB BULATS test version. The simulator also has a number of wider uses such as allowing the effects of amendments to algorithm parameters to be assessed, or even testing the effects of proposed changes to the algorithm itself, and also in informing the design of research projects by enabling recommended sample sizes to be estimated. The CAT Simulation Tool has therefore become an essential part of the CB BULATS test production toolkit.

References and further reading

Beeston, S (2000) The UCLES EFL Item Banking System, *Research Notes* 2, 8–9.

Flaugher, R (2000) Item Pools, in Wainer, H (Ed.) *Computerized Adaptive Testing: A Primer*, New Jersey: Lawrence Erlbaum Associates, 37–58.

Luecht, R, de Champlain, A and Nungester, R (1998) Maintaining Content Validity in Computerized Adaptive Testing, *Advances in Health Sciences Education* 3, 29–41.

Marshall, H (2006) The Cambridge ESOL Item Banking System, *Research Notes* 23, 3–5.

Thissen, D and Mislevy, R (2000) Testing Algorithms, in Wainer, H (Ed.) *Computerized Adaptive Testing: A Primer*, New Jersey: Lawrence Erlbaum Associates, 101–132.

Wainer, H (2000) *Computerized Adaptive Testing: A Primer*, New Jersey: Lawrence Erlbaum Associates

The comparability of computer-based and paper-based tests: goals, approaches, and a review of research

NEIL JONES AND LOUISE MAYCOCK RESEARCH AND VALIDATION GROUP

Introduction

This paper presents a position on how to approach the comparability of computer-based (CB) and paper-based (PB) test formats, as Cambridge ESOL prepares to roll out a new series of CB tests, including a computer-based version of the Business English Certificate (BEC) at Preliminary and Vantage levels. What are the practical goals of comparability, and what messages should we give to help people choose between formats or use them side by side? How should we conceptualise fairness when there is a choice between formats? What has research shown so far?

Technology enables things to be done more quickly, efficiently and conveniently. This is reason enough to adopt technologically-enhanced approaches to testing. Currently in the US, for example, the No Child Left Behind (NCLB) legislation is leading to a great expansion in the amount of testing going on in schools; and much of this testing is shifting towards being computer-based. In such situations a computer-based (CB) test is intended as a like-for-like replacement for, or alternative to, a paper-based (PB) test.

Additionally, technology may enable things to be done differently and better. CB tests have potential to improve on PB tests in quite radical ways. They might offer tasks which reflect more authentically the 'real world' skills we wish to test. Recording behaviour in detail, they might capture the processes students engage in while answering questions. Thus they might go beyond eliciting right and wrong responses to reveal the nature of the learner's current understanding or misunderstanding of a problem, in a way that can feed into learning (Pellegrino, Chodowsky and Glaser 2001). It is even possible to imagine traditional assessment disappearing completely within a technologically advanced learning environment which unobtrusively monitors and guides students (Bennett 1998, Jones 2006).

Cambridge ESOL has included CB tests in its portfolio of products since 1995, and is currently expanding the range of tests available in a CB format, so that by the end of 2007 CB versions will be available for all the following: IELTS, BULATS, KET, PET, BEC Preliminary and Vantage, ESOL Skills for Life and TKT. All of these will be made available alongside current PB versions. It is important for Cambridge ESOL to define an approach to comparability which will guide the validation of these particular assessments, while providing a more general framework for thinking about comparability of technology-based and traditional assessment. This is necessary if we are to give candidates and end users the information they need in order to choose between test formats and interpret results.

The goal of comparability studies: to inform users

There are four conclusions we might come to regarding the comparability of CB and PB formats in a particular situation:

1. That they can be used entirely interchangeably, i.e. it is 'a matter of indifference to the examinee whether the test is administered on computer or paper' (Bennett 2003:i).
2. That they differ to some extent for practical reasons inherent in the formats.
3. That they differ to some extent by design, in terms of the construct (so that one may be considered the better for some purpose).
4. That they differ so radically in terms of construct or purpose that comparison is not appropriate.

Case 1 offers a strong assurance of equivalence which may be difficult to provide. There is a class of issues where the CB user interface has been found to introduce construct-irrelevant sources of difficulty, although it may be a realistic goal to reduce or eliminate these by design improvements. Extended reading is an example discussed below.

Case 2 is thus a likely scenario. Practicality is one of the four crucial aspects of test design identified by Bachman and Palmer (1996). The other three aspects – validity, reliability and impact – must be optimised within the practical constraints of time, money, available technology etc. which are always present. While such constraints cannot be avoided, a problem arises if CB and PB formats of the same test are used for the same purpose, but operate under slightly different constraints. It may be impossible to assure strict equivalence of results. Given that there might be good practical reasons for using both formats, our responsibility is to provide information on any known effects to assist test users in interpreting results.

Case 3, where CB and PB formats test different constructs, has so far not been a major issue for Cambridge ESOL's CBT development, which has focussed mainly on producing CB analogues of PB task types, and exploiting tasks and statistical information (item difficulties) from existing PB item banks. It was important to do this to carry over to CB tests the interpretative frame of reference that had taken many years of effort to develop for the PB exams. However, as development turns towards exploiting the specific strengths of CB testing then the nature of the language proficiency construct tested will change. Will such CB tests be offered as alternative formats of current PB tests, or as different products? In any case, the differences in the constructs will have to be explained to users so that they can choose the test most suited to their purpose.

Comparability and fairness

The discussion so far suggests that maximizing comparability and minimizing format-specific effects is the best guarantee of fairness; however, this is to overlook the fact that comparability has a human dimension too. There is a largely generational gap between those who prefer to write on computers and those who prefer to write on paper. This profound difference undermines a traditional principle of assessment: that fairness is ensured by carefully standardising procedures. Imposing a standardised approach to the testing of extended writing may be held to effectively advantage or disadvantage a proportion of candidates. Standardisation is thus no longer a guarantee of fairness. Offering alternative CB and PB formats of a test may thus be seen in some circumstances as enhancing fairness and validity.

There is therefore a strong argument to be made that the traditional formulation of the comparability question – that it should be 'a matter of indifference to the examinee whether the test is administered on computer or paper' (Bennett 2003:i)– is outdated, at least for tests of certain skills. Generally, we should agree that fairness will be best served by ensuring that candidates can choose the test format which they believe allows them to demonstrate their ability to the full – a 'bias for best' approach. Adopting such an approach does not free us from the responsibility of studying format-specific effects, trying to remove or mitigate them where possible and desirable, and reporting our research to end users. But it is part of the realistic view we need to take of the practical and conceptual limitations on comparability.

In the next section we discuss the nature of comparability as a function of the relatedness of tests. Then in the remainder of the paper we review the history of empirical comparability studies and the findings of one recent Cambridge ESOL study concerning CB and PB formats of IELTS.

Levels of comparability

Mislevy (1992) makes the point that differences in the construct of *what* is tested and *how* it is tested limit the extent to which tests can be compared in a purely statistical way. Mislevy describes four levels of linking, of which we mention three here:

Equating allows tests to be used interchangeably, but is possible only if two tests have been constructed from the same test specification to the same blueprint. *Calibration* can link two tests constructed from the same specification but to a different blueprint, which thus have different measurement characteristics. Relating this to the CB-PB comparison, equating should be possible where both formats administer the same range of tasks from the same item bank and in the same manner, e.g. as a linear test with a particular time limit. CB and PB BEC should be comparable in this way. *Calibration* would cover the situation where, for example, the same item bank is drawn from but the CB test is adaptive and the PB test linear. This is the case with CB and PB BULATS. These are both strong forms of statistical linking which allow *a priori* setting of comparable standards. Nonetheless, empirical studies are still necessary to see whether in practice there are construct-irrelevant, format-specific effects.

Projection is an option where constructs are differently specified – tests do not measure 'the same thing'. It aims at predicting learners' scores on one test from another. It might be exemplified by the way that the US testing agency ETS have linked the internet-based, computer-based and paper-based formats of TOEFL by providing tables of score comparisons. These are offered as a guide, with the caveat that 'differences in the tests can make it difficult to establish exact comparisons' (ETS 2005). In such a situation both construct-relevant and format-specific effects combine to complicate comparison, and ETS rightly encourage receiving institutions to set standards which work satisfactorily for their specific situation.

Empirical evidence

Historically, studies relating to comparability of test forms have taken the view that it should be possible for tests to be used interchangeably, with a focus on attempting to demonstrate that candidates obtain the same scores on CB and PB forms of the same examination, subject to measurement error. For example, Paek (2005) reviewed a large body of comparability research focused primarily on the American schools sector, concluding that 'in general, computer and paper versions of traditional multiple-choice tests are comparable across grades and academic subjects' (Paek 2005:17). Two particular areas where there tends to be less clarity are where tests involve extensive reading passages or composition of long essay-type responses.

Paek (ibid) notes that CB tests with extended reading passages have tended to appear more difficult than their PB counterparts and she suggests that this may be related to differences in the reading comprehension strategies that candidates employ. For example, on paper, candidates may highlight certain relevant lines of text or use visual awareness techniques to remember where particular information is located on the page. It is argued that the use of scrolling (as opposed to pagination) to navigate electronic text inhibits the use of this 'visual learning' because the sense of placement on the page is lost (Paek, 2005:18). To some extent, this is likely to be true, but as people become more used to navigating electronic text on the Internet by scrolling, perhaps reading strategies will adapt accordingly and this issue will become redundant. In fact, scrolling was found to be a more popular method of navigating through text than pagination in trials of the computer-based version of PET (Hacket, 2005). Paek concedes that, particularly with new tools such as electronic highlighters being introduced in order to facilitate interaction with electronic reading passages which more closely mimics the way candidates interact with text on paper, there is further promise in this area.

For Writing assessment there is concern, not only for the way that candidates may interact differently with the test on computer, but also for the way examiner marking may be influenced by the presentation of scripts. It has been argued that the increased legibility of typed scripts over handwritten scripts have effects on examiners' perception of the writing, though there is a lack of consensus over the

direction of this difference: some have argued that untidy, illegible responses which have obvious crossings-out and editing are marked more severely, while others have found that handwritten scripts are marked more leniently, and have suggested that this could be because errors are easier to locate in typed scripts (see Shaw, 2003, for a more detailed account). Examiner training and standardisation clearly has an important role to play in ensuring that such effects are minimised as far as possible, so as not to impact unfairly on candidate scores.

Many studies of the effects of test mode on Writing assessment have attempted to account for preference in test format and computer familiarity, by comparing scores of candidates taking PB and CB assessments while controlling for some general measure of ability. Studies of this nature conducted by Russell and his colleagues (Russell 1999, Russell et al 2003, Russell and Haney 1997, Russell and Plati 2002) provided evidence that tests requiring candidates to complete assessments comprising open-ended and essay type responses on paper tended to underestimate the performance of those more used to writing on computer. They recommended that candidates be given the option of writing in the medium of their choice, arguing that test conditions would then better reflect real-world practice, making for a more authentic and valid assessment.

> In essence, the notion of allowing each examinee to customise their testing experience, such that they are able to demonstrate their best performance, is consistent with the practice of providing accommodations to students with special needs. To be clear, we are not advocating that examinees be provided with access to any tool or test format such that the accommodation itself leads to a higher test score. Rather, we suggest that examinees should be able to customise the environment in which they perform the test so that the influence of factors irrelevant to the construct being measured is reduced. (Russell et al 2003:289)

Horkay et al (2006) endorsed this view, having found computer familiarity to significantly predict performance on an online Writing test after controlling for Writing performance on paper. They argued that, while substantial numbers of students write better on computer, and others on paper, conducting Writing assessment in a single delivery mode will underestimate performance for those candidates not given the opportunity to write in their preferred mode. They do, however, point out the inextricable link to the test construct: 'Do we want to know how well students write on paper, how well they write on computer, or how well they write in the mode of their choice?' (Horkay et al, 2006:36), and this is something that test providers and stakeholders must consider.

This argument is comparable with the 'bias for best' approach outlined above, and reflects Cambridge ESOL's position on computer based testing to date: examinations such as IELTS, BULATS and PET are available both on paper and on computer in order to allow candidates to select the mode which suits them best (see Blackhurst 2005). This will also be the case for the new assessments due to be offered on Cambridge Connect, our online test delivery engine, over the coming year (see Seddon 2005). In fact, IELTS currently goes one step further, providing candidates who have chosen to take the computer based test with the option of handwriting or typing their responses to the Writing component.

Cambridge ESOL have carried out a number of comparability studies in relation to the introduction of CB IELTS (Green and Maycock 2004, Maycock and Green 2005, Blackhurst 2005) and the computer adaptive CB BULATS test (Jones 2000). However, if we acknowledge the view that we should not expect candidates to obtain the same score regardless of test format, but accept that they are likely to maximise their performance in the format which suits them best, then studies of this kind become limited in the amount of information they contribute. Studies similar to those conducted by Russell et al (2003), comparing candidates taking PB and CB assessments while controlling for some measure of general ability, may prove more useful. This has recently been attempted with data from live CB IELTS administrations, and the evidence appears to support the view that performance is not affected by test format, once ability is accounted for. Band scores for candidates taking the PB and CB forms of the test were compared using Multivariate Analysis of Covariance (MANCOVA): no significant differences were found between the performance of the PB and CB groups on the Reading, Listening and Writing components, once Speaking score was controlled for (Speaking is the same for PB and CB tests, and was included in the analysis in an attempt to control for general proficiency). A separate Analysis of Covariance (ANCOVA) was then carried out on CB candidates only, to assess the effect of opting to complete the Writing component on paper or computer. Again, there were no significant differences in Writing score between those who typed and handwrote their essays, once Reading, Listening and Speaking band scores were controlled for. It seems that this may be a useful approach to adopt for comparability work on other CB products.

Conclusion

This article has outlined some of the issues in considering the comparability of tests, specifically with reference to exploiting new technologies in assessment, and highlights the challenges test providers face in this area. Providing candidates with the opportunity to be examined in the mode of their choice is clearly the best way to maximise fairness and may also improve test authenticity, which is an important aspect of validity. However, in order for stakeholders to be able to interpret results appropriately and to make comparisons between candidates who have taken the same examination in different test modes, comparability studies remain important. The focus of such studies has changed over the years, and will continue to do so as CB assessment develops its potential, particularly in the field of formative assessment. Hence, as Cambridge ESOL's portfolio of CB assessments expands over the coming months and years, the issue of comparability will remain an essential component of our research agenda.

References and further reading

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

Bennett, R E (1998) *Reinventing assessment: Speculations on the future of large-scale educational testing*, New Jersey: Educational Testing Service Policy Information Center, retrieved from www.ets.org/Media/Research/pdf/PICREINVENT.pdf

Bennett, R E (2003) *Online Assessment and the Comparability of Score Meaning*, Educational Testing Service Research Memorandum RM-03-05, retrieved from www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf

Blackhurst, A (2005) Listening, Reading and Writing on computer-based and paper-based versions of IELTS, *Research Notes* 21, 14–17.

ETS (2005) *TOEFL Internet-based Test Score Comparison Tables*, retrieved from: www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Comparison_Tables.pdf

Green, T and Maycock, L (2004) Computer-based IELTS and paper-based versions of IELTS, *Research Notes* 18, 3–6.

Hacket, E (2005) The development of a computer-based version of PET, *Research Notes* 22, 9–13.

Horkay, N, Bennett, R E, Allen, N, Kaplan, B and Yan, F (2006) Does it matter if I take my writing test on computer? An empirical study of mode effects, in NAEP, *Journal of Technology, Learning and Assessment*, 5, (2), retrieved from www.jtla.org

Jones, N (2000) BULATS: A case study comparing computer based and paper-and-pencil tests, *Research Notes* 3, 10–13.

Jones, N (2006) Assessment for learning: the challenge for an examination board, in *Excellence in Assessment*, 1, retrieved from www.assessnet.org.uk/

Maycock, L and Green, T (2005) The effects on performance of computer familiarity and attitudes towards CB IELTS, *Research Notes* 20, 3–8.

Mislevy, R J (1992) *Linking Educational Assessments Concepts, Issues, Methods, and Prospects*, Princeton: ETS.

Paek, P (2005) *Recent Trends in Comparability Studies*, PEM Research Report 05-05, Pearson Educational Measurement, August 2005, retrieved from www.pearsonedmeasurement.com/research/research.htm

Pellegrino, J W, Chodowsky, N and Glaser, R (Eds.) (2001) *Knowing what students know: The science and design of educational assessment*, Washington, DEC: National Academy Press.

Russell, M (1999) Testing on computers: A follow-up study comparing performance on computer and on paper, in *Education Policy Analysis Archives* 7 (20), retrieved from http://epaa.asu.edu/epaa/v7n20

Russell, M, Goldberg, A and O'Connor, K (2003) Computer-based Testing and Validity: a look back into the future, *Assessment in Education* 10, (3).

Russell, M and Haney, W (1997) Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil, *Education Policy Analysis Archives* 5 (3), retrieved from http://epaa.asu.edu/epaa/v5n3

Russell, M and Plati, T (2002) Does it matter what I write? Comparing performance on paper, computer and portable writing devices, *Current Issues in Education [Online]* 5, (4), retrieved from http://cie.ed.asu.edu/volume5/number4

Seddon, P (2005) An overview of computer-based testing, *Research Notes* 22, 8–9.

Shaw, S D (2003) Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts, *Research Notes* 11, 7–10.

Modelling facets of the assessment of Writing within an ESM environment

STUART SHAW RESEARCH AND VALIDATION GROUP

Introduction

It is becoming increasingly important for Cambridge ESOL to be able to provide evidence of quality control in the form of assessment reliability and validity to the outside world. Whilst this is well advanced for objective tests of English, it is less so for the performance tests. In addition to the concern for reliability, a recent focus of language testing research has been the multiple features (or 'facets') of examiners and candidates engaged in tests that may systematically impact on test performance, but may not always be of relevance to the construct of communicative writing ability (see Milanovic and Saville 1996, *Introduction*, for an overview of the facets involved in performance assessments). Bachman regards facets as 'an aspect of the measurement procedure which the test developer believes may affect test scores and hence needs to be investigated as part test development' (Bachman 2004:416); as such they constitute potential sources of measurement error in test performance. Although these facets are already a consideration in the design and monitoring of Cambridge ESOL Writing tests, understanding of their impact remains limited and they are not systematically accounted for in reported test scores.

With the emergence of new technology an opportunity exists to radically alter the nature of future writing assessment. Electronic Script Management (ESM), for example, potentially provides the rater performance data necessary to gather evidence in a timely manner, which is particularly important for reliability. ESM also has the potential to facilitate new marking models that will enable tighter control over current assessment quality and costs. ESM should not only provide for the capture and management of writing test data, but should also open the facets of the test event to investigation, allowing for the adjustment of candidate scores if and as necessary. Conceptualising the assessment setting in terms of facets offers the possibility for estimation of the influence of examiner and task characteristics on ability estimates in the assessment setting. Moreover, such an approach can reveal interactions between different facets of the assessment situation which may have a systematic influence on scores. Ideally, the development of new approaches to scoring

writing should allow Cambridge ESOL to investigate and address the behaviour of the individual examiner when confronted with particular candidate responses or particular writing tasks.

This article describes efforts to conceptualise the Cambridge ESOL Writing assessment setting as a workflow in terms of the facets of the setting within an ESM environment.

Writing assessment as a 'marking model workflow'

A 'marking model workflow' is a procedural approach for processing a candidate's response. At its very simplest, a candidate in response to an input task in the testing instrument provides a specified output and the output is the mark awarded to the candidate as an estimate of the ability which is being tested. So, the input to a marking model is a candidate's work (e.g. a script) and the intended output is marks of known, consistent accuracy and precision that are independent of the marker who marked the script ('script' is the traditional Cambridge ESOL word although the term being used for ESM is 'response set', covering CBT responses, audio recordings, etc.). Various marking and quality assurance processes take place between the input and output stages. This process constitutes a basic workflow which can be built into the organisation's standard procedures (and can be quality-assured).

In workflow terms, a *work item* is an instance of a workflow (e.g. the processing of a particular response such as a complete script, single question or part question) on which the actions of the workflow will be carried out. These actions or *work steps* are carried out/initiated by a user (a designated class of user such as Principal Examiner (PE), Team Leader (TL) or Assistant Examiner (AE)). A *role* defines a class of user that has access rights and other properties affecting how users in that class can interact with the workflow system. *Work introduction* can be thought of as the initial state for a work item in the workflow system. Responses are imported into the workflow system and located in a marking work queue (a place where a work item is stored awaiting a work step action to be carried out). *Workflow routing* is the logical linkage between two or more work steps such that the completion of one work step causes one or more other work steps to become enabled and the work items to be loaded into the work queue associated with the enabled work step(s). Figure 1 shows the relationship between these elements.

The potential for electronic data capture within a workflow in an ESM environment is of particular value in the following areas:

- workflow conceptualisation
- identification and manipulation of facets which constitute the data in the workflow
- subsequent movement and reconciliation of data within the workflow.

The marking model workflow: basic building blocks

Marking models can, for example, be built from basic types of marking: single; double; multiple; review; and gold standard. Figures 2–6 show the basic work and data flows (the types of marking can be integrated into full marking models).

With single marking, one rater records his or her marks and annotations, which are loaded into a data store for subsequent processing. With double marking, a response is independently marked by two raters, who both record their marks and any annotations. Both sets of marks and annotations are loaded into the data store for subsequent processing. A variant on the double marking model is the multiple marking model, i.e. multiple observations (n) of the same sample of performance.

With review marking, a candidate's response and the original rater's marks and annotations are presented to the reviewer – such as a TL – who enters their own marks and annotations for loading into the data store. Both sets of marks and annotations are stored for subsequent processing, though generally the reviewer's mark will take precedence. With gold standard seeding, responses with pre-agreed but secret marks are introduced into a rater's work queue at certain intervals. A *gold standard* script is a clean copy of a script previously marked by a group of PEs. A Principal Examiner would determine, in consensus with a small group of other PEs or senior TLs and in advance of marking, what the score should be on a sample number of scripts. These scripts would then be introduced on a periodic basis throughout the marking period for marking by AEs. The rater is unable to distinguish these gold standard responses from other responses and marks them as normal, his or her marks and annotations being loaded

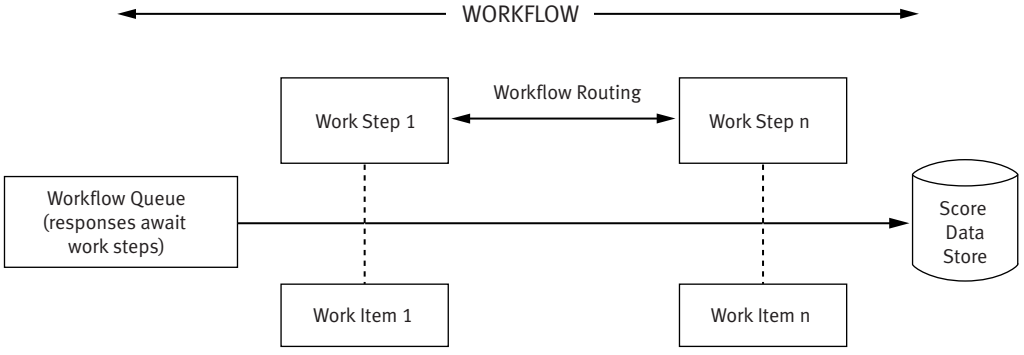

Figure 1: Elements of a marking model workflow

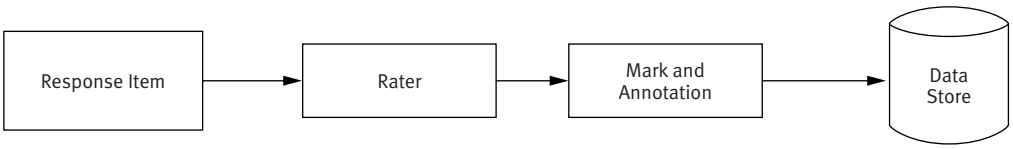

Figure 2:
Single marking model

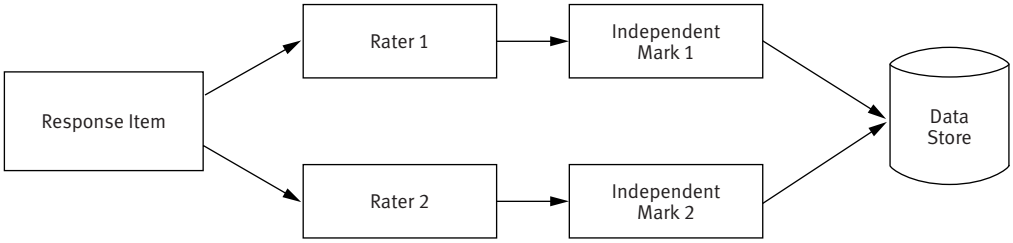

Figure 3:
Double marking model

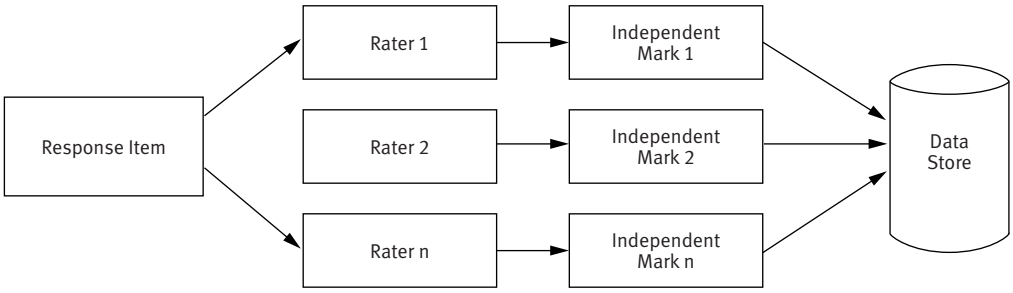

Figure 4:
Multiple marking model

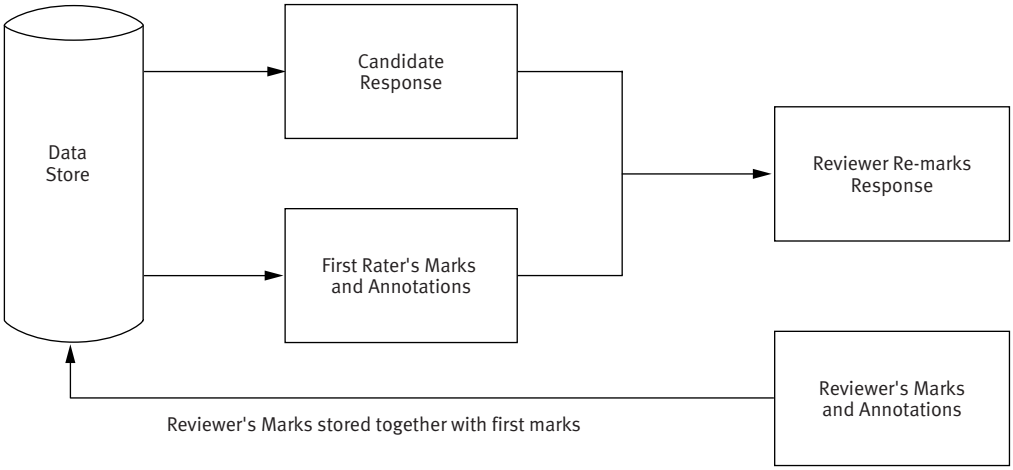

Figure 5:
Review marking model

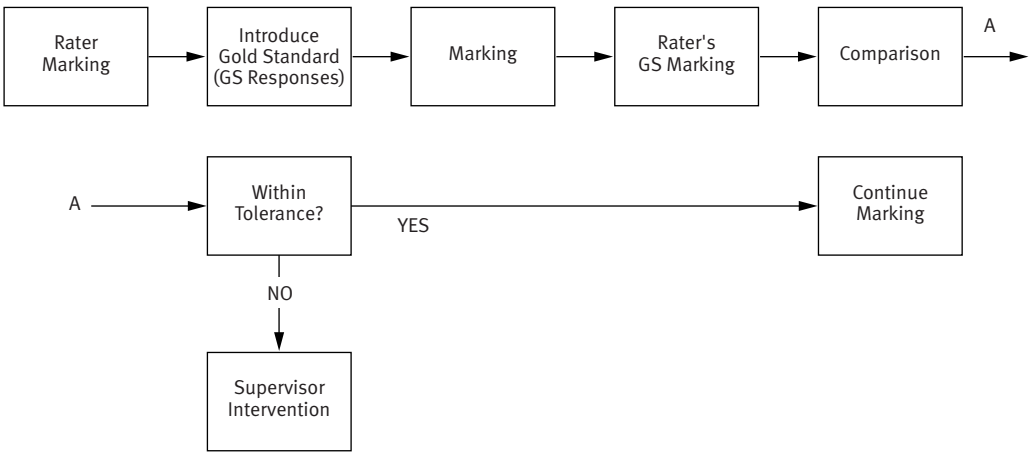

Figure 6:
Gold standard seeding marking model

into the data store for subsequent comparison with the gold standard marks. In this way a rater's marks may be compared with a gold standard that is independent of any particular TL or supervisor.

Conceptualising the writing assessment setting in terms of 'facets'

Writing tests are highly complex events involving multiple 'facets', some of which may be difficult to predict or control, which co-operate to produce a test score. Consideration of the various facets in writing assessment can reveal interactions between different facets of the assessment context which have a systematic influence on scores. The principal facets of the assessment context can be categorised into three groups: candidate (*ability*), task (*difficulty*) and rater (*severity*/*leniency*). Clearly, there are systematic facets of rater behaviour when confronted with particular candidates, particular tasks or particular test formats.

The modelling of the assessment characteristics – made possible by a faceted approach – has three primary functions:

1. A practical function – estimates of candidate ability may legislate for the features of both the rater and task thereby generating comparable candidate abilities which can be generalised across a universe of raters and tasks.
2. A planning and developmental function – for systems development, for example.
3. A research function – raising a host of research questions relating to facets of the rating assessment context.

A clearly articulated research agenda would be required in order to investigate the significance for specific measurement variables that any proposed model may suggest are likely to be of some importance. Examples of research questions relating to the facet of scoring might include:

- In what ways do raters differ? Is there a gender effect? (Facets of *Rater Status*, *Rater Profile* and *Rater Behaviour*.)
- Is it possible to identify distinct rater types and certain patterns of rater behaviour? (Facet of *Rating Behaviour*.)
- What amount of training/re-training is required? Can training improve raters' self-consistency? (Facets of *Rater Behaviour* and *Rating Training*.)
- How does assessment differ when marking electronically as opposed to paper-based marking? (Facets of *Rater Profile* and *Rater Behaviour*.)

Conceptualising writing assessment in terms of facets offers the potential for garnering complex data on the influence of rater and task characteristics on candidate ability estimates in the assessment setting (McNamara 1996). For example, it is conceivable that certain raters variably respond to candidates of particular L1 backgrounds or that gender effects may exist – where the gender of rater and candidate may influence scores (facet of *Rater Behaviour* or *Rater Profile*). It may be that the physical setting (which provides a context for the assessment) has an influence (facet of *Rater Setting*). In fact, any or all of the facets may exert a possible influence on the outcome of a test score. It is thus possible to collect information on the impact of any one of these facets (or any specific combination of them).

Each facet (or group of facets), assembled in a variety of ways, represents a potential source of data collection in the assessment context. If, for example, it is necessary to investigate finer-tuned aspects of the interaction of particular facets then the key facets must first be identified. These will constitute a focus for subsequent analyses. Facets of scoring validity, for example, can be identified and constructed for a particular assessment scenario, i.e. a particular kind of rating for a particular type of rater on a particular rating occasion. Suitable mechanisms for data collection and storage can be built into the workflow systems and it would be necessary to ensure that adequate data is both collected and stored for retrieval. Decisions as to whether data is required in real time for grading purposes or whether it is needed for subsequent validation purposes will need to be taken.

A facet approach enables the researcher to deconstruct any assessment setting into relevant constituent facets in order to address specific research questions relating to facets of the rating assessment context. In this way, facets can be assembled/re-assembled in a variety of different ways offering a number of key benefits:

- score matching through tasks to best reflect both the knowledge and ability of candidates, i.e. an effective scoring/procedural system
- knowledge and ability of candidates mediated through people (e.g. raters) and systems (e.g. scaling)
- introduction of stable and consistent scores
- ability to demonstrate an optimum marking model for score dependability
- greater control for assessment interactions
- introduction of control mechanisms through data collection.

An argument can be made for conceptualising facets of the assessment setting in terms of the various constituent validity parts of Weir's Socio-Cognitive Validation Framework (2005) described in *Research Notes* 21 (Weir and Shaw 2005). The framework has been developed with Cambridge ESOL and offers a perspective on the validity of Cambridge ESOL Writing tests (Shaw and Weir forthcoming). Of particular interest here, are the *a priori* validation components of context and cognitive validity and the *a posteriori* component of scoring validity (which together constitute what is frequently referred to as construct validity). Cambridge ESOL follows this socio-cognitive approach in relation to the Main Suite examinations where attention is paid to both context validity and to cognitive validity in terms of the cognitive processing and resources that are activated by test tasks. The 'superordinate' facets of *context*, *scoring validity* and the *test taker* can thus be deconstructed into sub-facets illustrated in Figures 7, 8 and 9.

Facet of Rating

- Automatic
- Semi-automatic
- Clerical
- Semi-skilled subject expert
- Expert

Facet of Rating Setting

- At home
- On-site
- Individual
- Group

Facet of Rating Profile

- Gender
- Experience
- Age
- Background
- Qualifications

Facet of Rating Timing

- Time of day
- Session
- Period of year
- Period into marking episode

Facet of Rating Training

- Recruitment
- Screening
- Induction
- Training
- Qualification
- Monitoring
- Grade review

Facet of Rating Status

- Expert (AE/TL/PE)
- Experienced
- Inexperienced
- New/Novice
- Subject expert
- Clerical

Facet of Rating Behaviour

- Severity
- Fit statistics
- Mean absolute difference from GS
- Skew
- Correlation
- Grade (last session)
- Duration

Figure 7: Facets of scoring validity

Task Data Facet

- Dichotomous
- Polytomous (without or ignoring judge meditation)
- Polytomous (taking judge meditation into account)

Response Facet

- Paper-based answer booklet
- Screen
- Multiple choice
- True/False
- Short Answer Question
- Extended Answer

Output Facet

- Transactional Letter
- Essay
- Report
- Discursive composition
- Short story
- Article
- Proposal
- Word completion
- Open cloze
- Information transfer
- Sentence transformations
- Review
- OCR/CIE task types

Channel Facet

- Receptive
- Productive

Delivery Facet

- On-demand
- Fixed date
- Paper-and-pencil
- Computer-based

Input Facet

- Instructions
- Rubric
- Task information

Figure 8: Facets of context validity

Candidate Skills

- Knowledge
- Experience
- Ability
- Use of computers
- World knowledge
- Cultural background

Candidate Profile

- Gender
- Age
- Nationality
- L1
- School type
- Years of study
- Exam preparation
- Exams taken
- Reasons for taking exam

Figure 9: Facets of the test taker

Conclusion

The search for a satisfactory conceptualisation of second language writing performance and for an adequate writing assessment model is a challenging one and it is clear that there is a need to broaden current discussions of the issues involved. It is hoped that the issues addressed here will make a positive contribution to the widening nature of the performance assessment debate within Cambridge ESOL, and within Cambridge Assessment more widely. Whilst the complexity and type of proposed assessment model have yet to be determined by research it is important that the model should be simple so that it is easily understood, easily implemented in software, and is computationally efficient – especially given the very large amounts of data that could be collected in an ESM environment. It is also a requirement that the model is able to identify problematic marking accurately and precisely, and that this can be improved by adding complexity, i.e. more sophisticated

modelling. In this sense, an iterative process of research and development is advocated, that starts with a simple model and subsequently adds complexity until the business is satisfied with the balance it has achieved. Such a model will probably explicitly model *candidate*, *task* and *rater* facets. There would be a need for the model to collect rich, robust data that facilitate the investigation and (if necessary) ongoing operationalisation of any issue considered relevant to understanding the nature of the assessment and promoting fairness. These data will need to embrace:

- details of the assessment (e.g. task parameters, administration, linguistic demands)
- candidate characteristics (e.g. gender, age, nationality, L1, years of study, knowledge, experience)
- examiner characteristics (e.g. qualifications, experience)
- rating details (clerical or 'expert' examiners, at home or residential rating, examiner behaviour, aspects of scoring).

Much of what is needed can already be found in existing examiner databases, test banks, candidate information sheet data stores and elsewhere. These sources would need to be related to test performances to facilitate investigations. As the scoring model is refined, some of this data might come to inform the operational management and calculation of test scores. The system would need to be designed with this level of flexibility in mind.

Capitalising fully on ESM would require a research agenda designed to:

1. Explore the significance for measurement of facets that the model may suggest are of importance (e.g. the effect of the characteristics of the rater on test scores).
2. Focus on the nature of interactions of the facets of assessment (e.g. the candidate/rater – rater/task interaction) especially given the interactional nature of performance assessment.
3. Ascertain what it is appropriate and realistic to both consider and assess in any given assessment context (position/stance adopted and a supporting rationale, feasibility and practicality of assessment proposals, etc.).

To meet the requirements of the Cambridge Assessment organisation as a whole, the outcomes of ESM would need to include a practical, practicable, feasible and manageable business system which builds on existing structures, but supports sufficient flexibility to accommodate the needs set out above and can also manage interactions between constantly changing systems (e.g. technical considerations, issues related to examiner payment and contracts). The harnessing of technology and corpus approaches to obtaining, storing and rating candidate performance is a major challenge that lies ahead for Cambridge ESOL.

References and further reading

Bachman, L F (2004) *Statistical Analyses for Language Assessment*, Cambridge: Cambridge University Press.

McNamara, T F (1996) *Measuring second language performance*, Harlow: Longman.

Milanovic, M and Saville, N (1996) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (Studies in Language Testing volume 3), Cambridge: UCLES/Cambridge University Press.

Shaw, S D and Weir, C J (forthcoming) *Examining Writing: Research and practice in assessing second language writing* (Studies in Language Testing volume 26), Cambridge: Cambridge ESOL/Cambridge University Press.

Weir, C J (2005) *Language Testing and Validation: An evidence-based approach*, Palgrave Macmillan.

Weir, C J and Shaw, S D (2005) Establishing the validity of Cambridge ESOL Writing tests: towards the implementation of a socio-cognitive model for test validation, *Research Notes* 21, 10–14.

Broadening the cultural context of examination materials

STEVE MURRAY ASSESSMENT AND OPERATIONS GROUP

Introduction

The following is a brief report on the Broadening the Cultural Context initiatives recently implemented within the context of Cambridge ESOL's production of material for its international examinations. The phrase *Broadening the Cultural Context* is used to refer to the importance of ensuring the cultural accessibility of our examination materials. It reflects a shared understanding that Cambridge ESOL's examination material should not contain a cultural focus which is too narrow, or which may favour the views or assumed knowledge of one culture over another.

For many years now Cambridge ESOL has produced sets of guidelines to assist Item Writers in their choice of materials for use in our assessments; these guidelines contain recommendations for topics and sources for materials, and also indicate that the materials should not reflect any kind of cultural bias. As a natural progression from this position, the Broadening the Cultural Context initiatives represent a development and an extension of current established practice. Following investigation and consultation, certain initiatives were identified and undertaken in order to ensure that materials continue to be fair and accessible to the widest proportion of the candidature. These initiatives included: the categorisation

of exam material in terms of the cultural context it could be considered to exemplify; the commissioning of ESOL consultants to author and conduct awareness raising activities for writers of our exam materials; and an evaluative stage, which aimed to gather data to describe any impact on the materials which the initiatives were felt to have achieved, from the perspective of the item writers.

This report begins by summarising the context and method of producing examination material at Cambridge ESOL. It goes on to describe the rationale, categorisation, training and impact of the initiatives and concludes with a brief comment on possible future directions.

The context and method of examination production at Cambridge ESOL

Cambridge ESOL offers a range of ESOL examinations to a worldwide candidature of around 1.8 million candidates in over 135 countries. Test material passes through four main stages in order to ensure a rigorous and quality driven approach to the production of examinations: commissioning and editing; pretesting; test construction; and question paper production. This final stage consists of the external vetting, checking and printing of the constructed papers or tasks. The stages of commissioning/editing, pretesting and test construction are the key stages during which the choice, consideration and editing of materials for assessment purposes is made. In terms of the commissioning of material, small teams of item writers, who may work on one or more of the components of the exams, are commissioned to source materials and write tasks according to the individual paper specifications. When these materials are submitted, they are initially evaluated at a pre-editing stage where they are accepted for editing or are rejected. Following editing, material is constructed into pretests to enable trialling on a representative sample of the international candidature; this ensures as far as possible that material and tasks are appropriate for the level and for the intended candidates. After completing the pretests, the candidates and teachers also have the opportunity to give written feedback on the topics and texts contained in them. When the materials are reviewed, candidate/teacher feedback on the materials is carefully considered alongside the statistical data which is gathered as a measure of its linguistic difficulty. Feedback which indicates material may not be suitable for any part of our international candidature informs the decision at this stage as to whether material progresses beyond this point into an actual examination paper. Material which meets the criteria, both statistically and in terms of appropriacy for the candidature, becomes available for test construction. At test construction, the constructed papers are carefully considered to ensure that, as well as fulfilling the statistical criteria at the level for which they are designed, each paper is accessible and appropriate for the situations, needs and wants of the target international candidature. Finally, there is an overview stage, at which the tests are considered alongside one another; this provides a global check on the overall appropriacy, quality and suitability of the examinations.

Rationale for broadening the cultural context

For some time, there had been a growing awareness that the increasing use of English as an international language, in which the target language use of candidates could be in a local or international context, was a need which language testers had a responsibility to ensure was appropriately reflected in testing materials. This perception was reinforced by data gathered by Cambridge ESOL on the purposes for which candidates were taking exams; such data indicated that candidates undertaking English-language examinations often do so for reasons other than work or study in the United Kingdom. It seemed there was some evidence that English-language test takers in our contemporary context may have no common need to engage with, visit or work in the culture from which their English-language test originated. Further impetus for the Broadening the Cultural Context initiatives arose partly out of a current focus in testing theory which stresses the importance of the appropriacy of examination materials for the target language use of the candidates, and partly from study of the data gathered by Cambridge ESOL's ongoing investigations into the needs and wants of the contemporary international candidature. All these considerations underpinned the rationale which led Cambridge ESOL to develop and implement a number of initiatives aimed at ensuring the continuing cultural accessibility of examination materials, and at raising the awareness of the issues among the item writer cadre. Additionally, and by no means a minor point when considering the cultural context of examination materials, it is an important shared assumption at Cambridge ESOL that in order to elicit the best performance from candidates in terms of their linguistic ability, examination materials ought, as far as possible, to be accessible, relevant and interesting to the target international candidature.

Preliminary considerations in broadening the cultural context

As a preliminary initiative in order to investigate the issues, Cambridge ESOL established a working group of key external consultants and internal staff who had roles in managing the production of examination materials. This group had the primary goal of identifying what principles or objectives could be said to underpin any subsequent initiatives in this context which might be developed and implemented within the examination production process. Following a series of consultations, this group identified, among others, the following objectives:

- It should not be assumed that candidates are knowledgeable about, or interested in, British culture.
- Texts and other task input material should not assume that candidates enjoy the lifestyles of particular income groups or nationalities.
- Exam materials should be drawn from as wide a range of sources as possible, including examples of non-British English.
- Good exam materials are often drawn from sources that have the widest possible original target audience.

- Where cultural assumptions might impede understanding, materials should be edited to gloss or, if necessary, remove cultural allusions or references.
- Assumptions should not be made about a candidate's location, cultural background and/or lifestyle.

With these objectives defined and agreed, two further initiatives were undertaken by Cambridge ESOL: the training of item writers, and the categorisation of materials.

Training of item writers

It was agreed that training should be rolled out to all item writers, partly as an awareness raising initiative regarding the cultural accessibility of exam materials, and partly to ensure that the issues were communicated and discussed as widely as possible. A number of experienced writers and consultants, including some of those from the original working group, were commissioned by Cambridge ESOL to author skill- and exam-specific training materials for the widest possible dissemination and use. In order to ensure these training materials were appropriate and useful for each of the contexts in which Cambridge ESOL offers international examinations, the training packs were tailored to particular exam contexts, including Business, Young Learner and General English. At their core, all training materials contained the exemplification of the fundamental principles for broadening the cultural context, which were identified and developed by the initial working group, and which it had been agreed should underpin any subsequent initiatives.

To assist the item writers, two sets of guidelines were developed using external and internal resources. These guidelines were designed to be companion documents to the Broadening the Cultural Context training, as a source of further reference for item writers. One document arose out of requests from item writers for assistance with technical issues associated with using the internet as a materials resource. The second document was written to offer writers assistance and guidance in their search for materials with a broader cultural context. This document draws on both published materials and the internet as potential sources, discusses some of the potential sources available, and contains sample texts which are commented on by experienced item writers and chairs in terms of their cultural perspective and suitability for use as examination material.

Categorisation of materials

Following a process of consultation and consideration, certain broad categories under which material could be classified were agreed; these would be used for a monitoring process to ensure an appropriate variety of material was progressing onto the examinations. It was considered important that the criteria for categorising materials should not be superficial, i.e. simply the names used, or the topic or location dealt with in a text; instead, the classification should be made at a fundamental level which considered the cultural context or overall perspective exemplified by any particular piece of examination material. In practice, it was decided that each item writer prior to the submission of material to Cambridge ESOL should consider and identify, according to the agreed broad criteria, the cultural context exemplified in each case. When material is submitted for consideration, according to ongoing Cambridge ESOL policy, material which is felt to exhibit cultural bias or too narrow a cultural context does not proceed beyond the pre-editing stage. At the other end of the production cycle, it was agreed that material should be checked at the overview stage, which reviews the complete exams prior to their going forward to external vetting and printing. These categorisations aim to ensure that all new material is classified according to the cultural context it exemplifies, and that all existing material which currently goes into tests would be subject to the same scrutiny.

Impact of training and categorisation

In terms of the impact of the categorisation initiatives at pre-editing and overview stages, a positive observation has been that item writers' awareness of issues such as the cultural accessibility of materials has been raised; the materials submitted to Cambridge ESOL at the pre-editing stage are observed to exhibit a broad range of widely accessible cultural contexts. The implications for future tests may also be positive, in that Cambridge ESOL can feel confident that its item writers are generating materials which will be, as far as possible, accessible and appropriate for the wide variety of international contexts in which its tests are taken.

In order to assess the impact of the training initiatives from the perspective of item writers, a survey was designed and administered by email to a representative sample of Cambridge ESOL's item writing cadre. The representative sample selection was obtained by a systematic method of selecting every third member of each item-writing team. The responses from the sample group indicated the following:

- The objectives for Broadening the Cultural Context were perceived by the item writers to be an appropriate and useful articulation of the frame of reference.
- The training and supporting documents were felt to be, on the whole, effective, adequate and useful.
- The training, which aimed to raise the awareness of issues of the cultural context of examination materials, has been perceived positively by the item writers and is felt to have had a positive impact on the materials.
- The training was felt to have been applied consistently and logically across the areas, such as General and Business English, in which Cambridge ESOL offers examinations.

The comments with which some of the sample group supported their answers were, in general, quite positive, for example: 'I think the training helped to dispel some myths and made the whole area more objective and concrete'. Finally, in the majority view of the sample group, it was considered that the training initiatives were appropriate to the contexts of the current candidature, and as a corollary that the training initiatives aimed at raising the awareness of issues of cultural accessibility and fairness were having and should continue to have a positive impact on Cambridge ESOL's international examinations.

Future directions for broadening the cultural context

Currently, although it is felt by Cambridge ESOL that the Broadening the Cultural Context initiatives have had a positive impact in increasing the cultural accessibility, relevance and so, it is hoped, the appeal of the examination materials, it is understood that efforts will continue to be ongoing in order to ensure the content of the exams remains appropriate and relevant with regard to the international candidature. Even as new item writers will receive training in the importance of the cultural accessibility and relevance of materials, efforts are continuing at all stages of the production of examination materials to ensure the cultural accessibility and appropriacy of materials. The overall aim of the approach is to ensure that each test has a balanced variety of materials, thus enabling the international candidature to come to all the tests as equally as possible in terms of content and cultural perspective. Over time, further initiatives in this and other areas may be identified and developed should the nature of the candidature change.

Research and development update

Asset Languages Update

The Asset Languages Research and Validation (R & V) team held a cross-language standard-setting day at Homerton College in Cambridge on the 30th September 2006. Examiners and moderators across twenty-two languages were represented at the day which used English exemplars in Speaking and Writing, covering Breakthrough to Advanced stages (approximately A1-B2), to set the standard.

The area of comparability is an important ongoing research focus for the R & V team. One area of current work with external consultants is developing a learner-centered model of comparability for reading. The model will cover three aspects: 1) *Texts* – analysing the features of reading texts, 2) *Tasks* – a close analysis of what the learner has to *do* and *know* for each test item; and 3) *Relating tasks to the real world* – what do the analyses from 1) and 2) tell us about the level of the task and learners who are successful on such tasks in Common European Framework of Reference (CEFR) terms? Once fully developed, this model will be used to analyse cross-language comparability in Asset Languages Reading papers. A similar model will also be developed for Listening.

As detailed by Ashton (2006), a Can Do self-assessment survey was developed and piloted for Reading. This survey has been used to compare learners' reading ability across languages (Urdu, German and Japanese). Similar surveys have now been developed in Listening, Speaking and Writing and will be piloted with groups of secondary and adult learners in the UK. As well as being used for cross-language comparability research projects, these surveys will be used to research learners' relative ability across skills. This information will be looked at with item level analysis from the same learners to help ensure that standards set across the four skills are comparable.

Grade comparability between the two strands of Asset Languages (Teacher Assessment and External Assessment) is also currently being researched. Although work in this area is ongoing, analysis to date shows that there is good comparability between Teacher Assessment and External Assessment grades for Listening, Reading and Speaking. There is less comparability for writing, with teachers awarding slightly higher grades than were achieved in the External Assessment. As there were only a small number of students in this sample though, further data is needed to determine whether this is a general trend.

For more information on Asset Languages visit www.assetlanguages.org.uk

References and further reading

Ashton, K (2006) Can do self-assessment: investigating cross-language comparability in reading, *Research Notes* 24, 10–14.

ESOL Special Circumstances

The work of ESOL Special Circumstances covers three main areas: special arrangements, special consideration, and cases of malpractice in respect of ESOL products. Special arrangements are made for candidates with a permanent or temporary disability, e.g. learning, hearing, visual or motor difficulties, to enable them, as far as possible, to take the examination on an equal footing with other candidates. Special consideration is given to candidates who are affected before or during an examination by adverse circumstances, e.g. illness, bereavement, unexpected interruption or excessive noise. Malpractice concerns any conduct which has the intention or effect of giving unfair advantage to one or more candidates; cases are brought to the attention of Cambridge ESOL via reports from centres, reports from examiners or inspectors, and through routine statistical checks applied to candidates' answers.

A Special Circumstances Report is prepared on an annual basis and provides a general survey of work carried out, together with an analysis of cases dealt with. The most recent annual report, for 2005, is now available via the Cambridge ESOL website at:
http://www.cambridgeesol.org/research/special.htm

Conference reports

Language Testing Forum – November 2006

This year's Language Testing Forum was held at the University of Reading and was hosted by the Centre for Applied Language Studies (CALS). The introductory lecture entitled *Spoken fluency – theory and practice* was given by Prof Mike McCarthy (Nottingham University). Common perceptions of fluency were considered and questioned in light of corpus evidence which has shown that native speakers are often *dysfluent* and uneven in their spoken performance. After listening to a dialogue between a native and non-native speaker the group discussed the fluency and relative contribution of each speaker. The idea of conversations rather than speakers being fluent was suggested, in that speakers communicate strategically and cooperatively and are responsible for the continuation of the conversation. The implications for oral assessment were discussed, particularly the role of interlocutor.

Two papers were presented by members of Cambridge ESOL. Martin Robinson described the role of content specialists in both the development and production of ESP tests, in particular the International Legal English Certificate (ILEC) and the International Certificate of Financial English (ICFE). In order to deliver high quality testing instruments, it is vital that there is a full understanding of the language use domain and the degree of task authenticity. When testing for specific purposes or contexts, experts may be enlisted to provide some of this information. Traditionally, content specialists have been used only at the needs analysis stage of test development; however, in the development of ILEC and ICFE these specialists have been used throughout the design and development process in an integrated approach. In addition, this collaboration is continuing with ILEC through out the item and test production process. Martin outlined the issues involved with this approach and described the nature of working with specialist groups.

Ardeshir Geranpayeh detailed the use of Structural Equation Modelling (SEM) to inform the revision of FCE. Ardeshir began with a history of FCE test development and revision before outlining some of the empirical procedures used in the current revision. When revising any test it is essential that any changes in format do not significantly alter the underlying test constructs; Ardeshir described two studies in which the construct of the FCE exam were explored. The first involved the production of a number of construct models, based on data from one major FCE examination session. The viability of each model was then tested using SEM and confirmatory factor analysis (CFA). A model which best described the underlying constructs was chosen, and then verified in a second study using data from two further FCE sessions. This model, which is based on a componential view of language testing, is now being used to influence the current revision process.

There were a variety of interesting papers and posters given by delegates from a number of institutions including Lancaster University, the University of Bristol, Roehampton University, the University of Essex and the University of Reading. Topics covered included: language support in EAL classrooms, effect of task on group speaking tests, reading to summarize and the TOEFL impact study. In addition Prof Alan Davies (University of Edinburgh) led an open discussion on the ILTA draft Code of Practice (CoP). In this workshop participants were invited to discuss whether it is necessary for organisations such as ILTA to have a CoP in addition to a Code of Ethics (CoE) and to comment on the draft CoP which has recently been developed. Issues discussed included whether it would be more appropriate to have an expanded CoE, whether a CoP should be mandatory or voluntary, and who the intended audience would be.

IELTS Masters Award

Winner of IELTS Masters Award 2006

In 1999, the three IELTS partners – the University of Cambridge ESOL Examinations, The British Council, and IDP: IELTS Australia – inaugurated the IELTS MA Thesis Award, an annual award of £1000 for the masters level thesis or dissertation in English which makes the most significant contribution to the field of language testing. Since 1999, there have been 5 winners of the award – from Canada, Australia, USA and the UK.

For the 2006 IELTS Masters Award, submissions were accepted for masters theses completed and approved in 2005. The IELTS Research Committee, which comprises members of the three partner organisations, met in November 2006 to review the shortlisted submissions and the Committee was once again impressed with the quality of work received. After careful consideration, the Committee decided to announce one winner: Youn-Hee Kim – for her thesis entitled 'An investigation into variability of tasks and teacher-judges in second language oral performance assessment'. Youn-Hee completed her thesis at the Department of Integrated Studies in Education, McGill University, Montreal (Canada) and her supervisor was Dr Carolyn E Turner.

Youn-Hee's full abstract appears below:

While performance assessment has broadened and enriched the practice of language testing, ongoing questions have arisen as to whether complexity and variability in performance assessment influence a test's usefulness. That testing tools and human factors must be involved in test-taking and rating procedures is inevitable, but these factors have long been recognized as potential sources of variance that is irrelevant to a test's construct. This study continues the ongoing discussion about rater and task variability by comprehensively examining how second language oral performance is assessed by different groups of teacher-judges across different task types. The substantive focus of the study investigated whether native English-speaking (NS) and non-native English-speaking (NNS) teacher-judges exhibited internal consistency and interchangeable severity, and how they influenced task difficulty and the calibration of rating scales across different task types. It was also identified what the salient construct elements for evaluation were to the two groups of teacher-judges across different task types when no evaluation criteria were available for them to consult.

A Many-faceted Rasch Measurement analysis of 1,727 ratings and a grounded theory analysis of 3,295 written comments on students' oral English performance showed little difference between the NS and NNS groups in terms of internal consistency and severity. Additionally, the two groups were neither positively nor negatively biased toward a particular task type. The NS and NNS groups, however, did differ in how they influenced the calibration of rating scales, and in how they drew upon salient construct elements across different task types. The suitability of the NNS teacher-judges, the need for context (task)-specific assessment, the usefulness of the Many-faceted Rasch Measurement, and the legitimacy of mixed methods research are discussed based on these findings.

The Research Committee noted that Youn-Hee's dissertation was work of a very high standard. The research was well grounded within the literature and was both thorough and pertinent. Her research questions were well focused and clearly measurable. Multi-faceted Rasch was skillfully used and interpretation of results clear and perceptive. Any conclusions were supported by evidence garnered from the work. This was an important piece of research, of considerable interest to language testing specialists and was a worthy winner of the 2006 award.

Youn-Hee will be presented with her award (a cheque and certificate) at the 29th Annual Language Testing Research Colloquium (LTRC), *9–11 June, 2007* at the University of Barcelona, Spain. For further information about LTRC see www.iltaonline.com/LTRC_07_Announce.htm.

Call for Entries for IELTS Masters Award 2007

Each year the IELTS partners 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. The entry procedures and timetable for the 2007 award are given below.

Submission and evaluation procedures

Dissertations will only be considered eligible if they were submitted and approved by your university in 2006. Dissertations completed in 2007 will not be considered eligible for the 2007 award but may be submitted the following year. Submissions should be for dissertations written in partial or total fulfilment of the requirements for a Masters degree or its equivalent.

The full dissertation abstract, accompanied by both the *Introduction* and *Method* chapters together with a reference from your supervisor, should be submitted to:

Dr Lynda Taylor/Stuart Shaw
Research and Validation Group
University of Cambridge ESOL Examinations
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

The IELTS Research Committee will review the submissions and shortlist potential award winners. For all shortlisted dissertations a full copy of the dissertation will be requested and a further reference may be sought. Shortlisted dissertations will be reviewed and evaluated by the IELTS Research Committee according to the following criteria:

- rationale for the research
- contextualisation within the literature
- feasibility of outcomes
- design of research question(s)
- choice and use of methodology
- interpretation and conclusions
- quality of presentation
- use of references
- contribution to the field
- potential for future publication.

The Committee's decision is final.

Timetable

The following timetable will apply in 2007:

1 June Deadline for submission of dissertation extracts and supervisor's reference to Cambridge ESOL.

1 August Deadline for submission of full copies of shortlisted dissertations (and further references if required).

October/**November** Meeting of IELTS Research Committee.

November/**December** Announcement of award.

Please note that submission details may change from year to year and it is therefore important that the most current procedures are consulted. Details of the application process for the IELTS Masters Award 2007 can also be found on the IELTS website: www.ielts.org