ResearchNotes

Contents

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

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

Published February 2004

Editorial Notes

Welcome to issue 15 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

This issue focuses on aspects of test development and revision for our specialised language testing products: those that test business (BEC) and academic English (IELTS); those aimed at young learners (YLE) and modular tests (CELS). Although our Main Suite exams are the oldest and probably best known set of tests we offer, our specialised tests are equally important in a rapidly changing world. In this issue we provide insights into the range of test development activities that we undertake including revising tests and their underlying constructs, collecting evidence of the impact of our exams, and training examiners. We also have a new section which presents performance data for recent exam sessions.

In the opening article Lynda Taylor considers test comparability, an important consideration for anyone involved in language testing. She discusses score and test equivalence then describes the pros and cons of comparative frameworks before summarising what Cambridge ESOL is doing in this area.

In the following article David Horner and Peter Strutt (University of London British Institute in Paris) report recent research into the productive vocabulary of business English students in which they explore lexical categories with native and non-native informants using BEC data from the Cambridge Learner Corpus. These findings could help inform task design and assessment criteria for BEC. Continuing this theme, Stuart Shaw describes the latest phase of the IELTS Writing assessment revision in which he explores underlength and illegible scripts and task rubrics, and suggests what validation work is required.

The next pair of articles describe the impact of specialised exams. Roger Hawkey summarises Phase 3 of the IELTS Impact Study which seeks to monitor the impact of IELTS world-wide, presenting selected findings on the impact of IELTS on preparation courses and candidates' views of IELTS. On a smaller scale, Trish Burrow and Juliet Wilson report on a recent stakeholder survey which forms part of the ongoing review of YLE, our suite of language tests for children.

Moving to rater training, Stuart Shaw describes how we are creating a virtual community of raters using on-line technology, with a view to improving the reliability of marking writing. This is being trialled with CELS writing examiners although it has relevance for all of our exams. Ardeshir Geranpayeh continues the discussion of reliability, presenting the view from the language testing literature and reliability estimates for FCE objective papers from 2000–2003.

We end this issue with the announcement of the winner of the IELTS Master's Award 2003 and details of how to enter for the 2004 award.

Issues of test comparability

LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

Introduction

What's the difference? is a question which is increasingly asked in the world of language testing. *What's the difference between Test A and Test B? How do they compare?*

The ability to relate different tests to one another in useful and meaningful ways is becoming more and more important for test users. As a result, test providers are being challenged to pay greater attention to issues of test comparability – both in terms of the relationship between different tests within their own product range (in Cambridge ESOL's case, for example, *What's the difference between First Certificate and BEC Vantage?*), and also in terms of the relationship between their own assessment products and those offered by competitor boards (for example, *What's the difference between IELTS and TOEFL?*)

There have probably always been informal attempts to compare language proficiency measures; traditionally, comparisons have tended to focus on the notion of 'score equivalences', i.e. how do the scores or grades from two different tests relate to one another, and to what extent can they be considered 'equivalent'. A more formal attempt took place in 1987 when Cambridge was involved in the 3-year Cambridge–TOEFL Comparability Study, set up at the instigation of UCLES and carried out under the direction of Professor Lyle Bachman; an account of this ambitious and innovative study was published as the first volume in the Studies in Language Testing series (1995). In the preface to the volume, Bachman reminds readers that any comparability study needs to take account of more than just score equivalences; it must also investigate comparability of test content and performance.

Defining our terms

It may help to define terms such as 'equivalence' and 'comparability' in the language testing context. 'Equivalence' is often used with reference to 'equivalent forms' for which the ALTE *Multilingual Glossary of Language Testing Terms* (1998) offers the following definition:

> *Different versions of the same test, which are regarded as equivalent to each other in that they are based on the same specifications and measure the same competence. To meet the strict requirements of equivalence under classical test theory, different forms of a test must have the same mean difficulty, variance, and co-variance, when administered to the same persons.*

Cambridge ESOL produces different versions – also known as 'alternate' or 'parallel' forms – of the same test to be taken on different session dates throughout the year; tests must clearly be equivalent from session to session in terms of their content coverage and measurement characteristics. The Glossary notes that "equivalence is very difficult to achieve in practice" and it's fair to say that considerable effort and expertise goes into ensuring test equivalence through the implementation of a comprehensive set of standard procedures applied at each stage of test production.

Under its entry for the term 'test equivalence', the *Dictionary of Language Testing* by Davies et al. (1999) offers a similar definition to the one given above and goes on to mention the increasingly common use of IRT analysis and item banking to help with the process of creating equivalent forms, an approach which has become fundamental to the Cambridge ESOL test production process.

However, the dictionary offers a further definition of the term 'equivalence': *the relationship between two tests*; it goes on to add:

> *Strictly speaking, this concept is unjustifiable, since each test is designed for a different purpose and a different population, and may view and assess language traits in different ways as well as describing test-taker performance differently.*

This should be good news for test producers for if the concept of 'test equivalence' across two different tests is strictly speaking unjustifiable, perhaps we don't need to be concerned with it. But the dictionary entry continues: *However, in reality test users may demand statements of equivalence between different tests (for example, admissions officers at educational institutions).*

Of course this is exactly what happens. But it's not surprising that test users seek firm statements concerning the 'equivalence' or otherwise of different tests. University admissions officers want to know how to deal with students who present them with TOEFL, IELTS or CPE scores; employers need to know how to interpret different language qualifications previously achieved by potential employees; schools, teachers and students have to make choices about which test to take and they want to be clear about the relative merits of those on offer (e.g. FCE or BEC Vantage or CELS Vantage). All this is set against a socio-economic backdrop of rising mobility of the population, increasing consumer choice and a growing emphasis on accountability. Today's world is characterised by an overwhelming range of options from which to make your selection. In a similar way, learners, test takers and other test users have become consumers with rights and expectations; one expectation is that test providers will provide as much useful information about their products as possible to assist consumers in their purchasing decisions, and this is exactly what as a responsible examination board we seek to do using a range of methods including: printed publications, electronic information on our global and in-country websites, in-country teacher seminar programs, etc.

Identifying features for comparison

It is clearly possible to make comparisons between or across different products claiming to measure language proficiency. But what are the essential criteria which should form the basis of any meaningful comparison, and how can the outcomes of a comparative analysis be reported in a useful and meaningful way? And to what extent are those criteria which language testers (designers or researchers) focus on when comparing tests the same criteria as those which test users (teachers, candidates) find relevant or salient when making comparisons?

When we set out to buy a washing machine or a portable stereo we generally seek information about criterial attributes of the product of interest – features such as price, dimensions, efficiency, ease of use, etc; using this information we compare features across different products and weigh up their relative importance in order to make our final purchasing decision. The information might be supplied to us by a salesperson in the retail trade and they may or may not be able to offer us an independent, objective analysis. Alternatively, we might consult a more independent consumer survey such as *The Good Shopping Guide* or *Which? Magazine*; both these publications identify a range of potentially relevant product features for comparison, and do the comparative analysis for you. An example from the June 2003 *Which? Magazine* illustrates how this works for disposable cameras. Analysis is based upon two dimensions: an objective 'specification', and an evaluative set of 'ratings'; each dimension covers a number of features or criterial attributes:

- Specification: price, number of exposures, film speed, waterproof
- Ratings (on a 5-point scale of best to worst): flash recovery time, sharpness of picture, picture quality (sunny + overcast), overall picture quality, ease of use

In addition, a total test score is calculated for each product which weights and combines the ratings for particular features to generate a final % score.

Is it possible to think of a language test like any other consumer product, e.g. a disposable camera or a portable stereo? Is it reasonable for a test user to be able to consult some sort of 'consumer survey' to inform their 'choice of purchase'? In fact, has *Which? Magazine* ever featured comparative analyses of tests or test instruments?

Interestingly, over the past year the magazine has twice focused on tests – although admittedly not language tests. In August 2003 it carried a report on 'genetic tests', and in November 2003 it profiled 'eye tests'. Both articles make fascinating reading for they reflect many of the concepts and much of the discourse that we are familiar with in the world of language testing.

The criteria the consumer analysts focused on when comparing the different genetic and eye tests currently available to the public were the following:

- Genetic tests: genetic risks, types of test, what tests can tell you, test availability
- Eye tests: accuracy of prescription/measurement/diagnosis, thoroughness, speed of testing, qualifications of test providers, interpretability of results, cost

The criteria listed here are not really so very different from some of the criterial features we might focus on when comparing different language tests: purpose, construct definition, test method, content breadth and depth, skills coverage, accuracy of measurement, predictive/diagnostic power, score interpretability, test length, accessibility, and cost. There are additional criterial features we might wish to add such as: degree of specificity, currency and recognition, relationship to curriculum, impact in the wider world. At Cambridge ESOL we have found it helpful to group this large collection of criterial features under the four overarching test qualities of:

- Validity
- Reliability
- Impact
- Practicality

This means our test design and development becomes a matter of seeking to achieve an appropriate balance among the four overarching qualities in order to try and ensure that any test achieves a sufficient level of usefulness, fairness and fitness for purpose.

In a similar way, Weir (forthcoming) highlights key elements which he believes underpin test fairness and which test developers need to address. In his forthcoming book on *Language Testing and Validity Evidence* Weir views commercial exams in the same way as products for sale; as such he considers that they should be subjected to similar scrutiny in relation to the following six high-level criteria:

- The test-taker
- Content validity
- Theory-based validity
- Scoring validity
- Consequential validity
- Criterion-related validity

Weir suggests that the six criteria can be used to generate key questions which test developers and users should ask of any test. He writes: *When we are buying a new car or a new house we have a whole list of questions we want to ask of the person selling it. Any failure to answer a question or an incomplete answer will leave doubts in our minds about buying. Poor performance in relation to one of these questions puts doubts in our minds about buying the house or car.* Similar questions which he suggests could be asked of a commercial test product are:

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test?
- Are the characteristics of the test task(s) and its administration fair to the candidates who are taking them?
- Are the cognitive processes required to complete the tasks appropriate?
- How far can we depend on the scores on the test?
- What effects does the test have on its various stakeholders?
- What external evidence is there outside of the test scores themselves that it is doing a good job?

So far the focus in this article has been on individual criterial features or attributes according to which tests can be compared and how these might be grouped into categories. In this final part of the paper we shall consider the growing role being played by comparative frameworks when comparing two, or more often, several different tests – sometimes across languages.

Comparative frameworks – benefits and risks

A focus on individual criteria – such as content, availability, or price – can deliver simple, but meaningful and useful comparisons between different tests. Comparative frameworks, however, are communicative tools which attempt a far more ambitious comparison of tests; and while they promise certain benefits they can also carry inherent risks. This is because all frameworks, by definition, seek to summarise and simplify, highlighting those features which are held in common across tests in order to provide a convenient point of reference for users and situations of use. Since the driving motivation behind them is usefulness or ease of interpretation, comparative frameworks cannot easily accommodate the multidimensional complexity of a thorough comparative analysis; the framework will focus on shared elements but may have to ignore significant differentiating features. The result is that while a framework can look elegant and convincing, it may fail to communicate some key differences between the elements co-located within it. The result is likely to be an oversimplication and may even encourage misinterpretation on the part of users about the relative merits or value of different exams.

In 1985 the English-Speaking Union (ESU) set up the ‘Framework’ project to try and devise a comprehensive frame of description to compare the various examinations of the main English Language Boards. The result was the ESU Framework (Carroll and West 1989) which covered 15 of the principal Examination Boards in English as an International Language and described exams in a standardised way, drawing on a series of performance scales. Its aim was to give users help in choosing the most suitable tests for them and interpreting the results.

Cambridge ESOL (then UCLES EFL) took part in the original exercise which attempted to map the exams on to a 9 band (ELTS-type) scale. The Cambridge exams therefore appeared on the first ESU chart. In the early 90s when the ESU wanted to revise the chart, Cambridge ESOL voiced a number of concerns over the presentation of information on the chart and in the end declined to take any further part in the project.

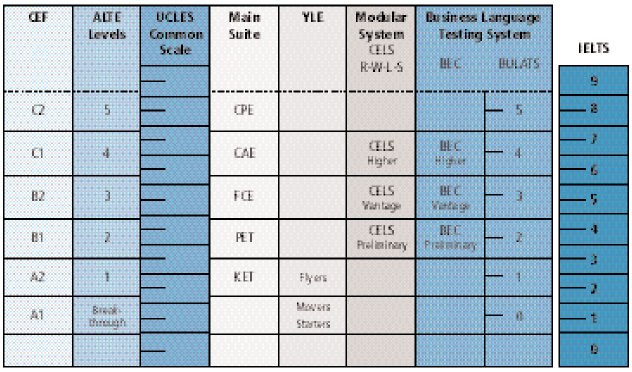
Staff in Cambridge were concerned about the way the chart was being used to claim “equivalence” and in particular believed that the two dimensional presentation of information on the Framework failed to reflect adequately a number of key differentiating features across the examinations described. There was concern that because the Framework implied greater degrees of comparability than was actually justified, it risked oversimplification or misinterpretation and this limited its usefulness to the users for whom it was intended. For example, it did not take into account the currency and recognition of the certificates nor the ability of the organisations represented to address issues of quality and fairness. A major concern was the degree to which exam boards were able to account for the standards of their exams in relation to the stability and consistency of their levels over time – a prerequisite for comparisons with any other exam.

The issue of comparative frameworks and their integrity is a particular challenging one for us in Cambridge. It’s well recognised that Cambridge ESOL has never subscribed to a philosophy of ‘one size fits all’ in relation to English language assessment. Over time, and in response to market demands and opportunities, we have developed a wide range of assessment products which now includes: tests at different proficiency levels (KET, FCE, CPE); tests involving a multi-skills package (IELTS) and tests which are modular (CELS); tests across different language domains (General English, Business English); tests for adult teachers of English and tests for young learners of English; tests in paper and pencil mode (Standard BULATS) and tests in computer mode (CB BULATS); tests for certificated use and tests for institutional use.

Clearly, the development of such a varied product range has led to us needing to build a frame of reference which explains to existing and potential test users the distinctive nature of each product we offer and how it relates to all other products in our range. For this reason work has gone on in recent years to explore the relationships between our different tests in an attempt to place each one within a wider framework of reference. As more tests were added to the Cambridge ESOL product range, so a conceptual framework of reference began to take shape in the 1990s – the Cambridge 5-Level System; this was later aligned with the developing ALTE 5-Level Framework which sought to recognise links between the different foreign language testing suites produced by European testing agencies such as the Alliance Française and the Goethe Institut. More recently, the ALTE Framework has been linked to the Council of Europe’s Common Framework of Reference (2001) which has both a theoretical and empirical dimension to it, and which has been gaining in recognition in the European educational domain. Strictly speaking the Common European Framework (CEF) is a framework of levels for the purpose of reference rather than test comparison. Its usefulness resides in its attempt to locate different tests according to broadly defined levels of proficiency. The CEF can tell us little, however, about the differences between tests which have been notionally located at the same level. For that we would need a multifaceted framework which asks (and answers) a range of questions about each test's characteristics.

In recent years we have found it helpful with both internal and external test stakeholders to communicate the relationship – in broad terms – between Cambridge ESOL tests through the conceptual framework of reference shown in Figure 1.

Figure 1: Conceptual framework of Cambridge English Language Testing Systems



CEF	ALTE Levels	UCLES Common Scale	Main Suite	YLE	Modular System CELS R-W-L-S	Business Language Testing System BEC	Business Language Testing System BULATS
C2	5		CPE				5
C1	4		CAE		CELS Higher	BEC Higher	4
B2	3		FCE		CELS Vantage	BEC Vantage	3
B1	2		PET		CELS Preliminary	BEC Preliminary	2
A2	1		KET	Flyers			1
A1	Break-through			Movers Starters			0

We are aware, however, that the process of placing tests on a common scale or within a common frame of reference is no easy matter. Conceptually it may be possible and even desirable to be able to co-locate different tests at shared proficiency levels (e.g. B2 on the CEF) or along common dimensions (e.g. social and tourist, study, work) but we always need to bear in mind that earlier 'health warning' in the language testing dictionary:

> *each test is designed for a different purpose and a different population, and may view and assess language traits in different ways as well as describing test-taker performance differently.*

The conceptual co-location of tests within a common frame of reference such as that in Figure 1 inevitably leads on to the challenge of providing empirical evidence to support it. Since the early 90s the Research and Validation Group has been using latent trait methods to link the levels of the Cambridge exams onto a common measurement scale. More recently, work has focused on defining a Common Scale for Writing through analysis of Writing performance across different proficiency levels and across different domains (as realised in the writing test scripts of Main Suite, BEC and IELTS test-takers). Investigations are also ongoing into issues of cross-language equating to determine what functional equivalence of ability means across different European languages. Clearly the conceptual framework shown in Figure 1 will continue to evolve as empirical evidence is gathered in support of links between levels and tests, and also as new, emerging test products need to be located within the larger frame of reference for Cambridge ESOL exams.

Conclusion

There is no doubt that comparative frameworks can serve a useful function for a wide variety of test stakeholders: for test users – such as admissions officers, employers, teachers, learners – frameworks make it easier to understand the range of assessment options available and help users to make appropriate choices for their needs; for applied linguists and language testers frameworks can help define a research agenda and identify research hypotheses for investigation; for test providers frameworks not only help with product definition and promotion, but also with planning for future test design and development. But we need to understand that they have their limitations too: they risk masking significant differentiating features, they tend to encourage oversimplification and misinterpretation, and there is always a danger that they are adopted as prescriptive rather than informative tools. They need to come with the appropriate health warnings!

References and further reading

ALTE (1998): *A Multilingual Glossary of Language Testing Terms, Studies in Language Testing*, Volume 6, UCLES/CUP.

Bachman, L F, Davidson, F, Ryan, C and Choi, I C (1995): *An investigation into the comparability of two tests of English as a foreign language – The Cambridge–TOEFL Comparability Study*, Studies in Language Testing, Volume 1, UCLES/CUP.

Carroll, B J and West, R (1989): *The ESU Framework Performance Scales for English Language Examinations*, London: Longman.

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

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T and McNamara T (1999): *Dictionary of Language Testing, Studies in Language Testing*, Volume 7, UCLES/CUP.

The Good Shopping Guide (2002) published by The Ethical Marketing Group.

Weir C (forthcoming): *Language Testing and Validity Evidence*, Basingstoke: Palgrave Macmillan Ltd.

Which? Magazine, published by the Consumers' Association.

Analysing domain-specific lexical categories: evidence from the BEC written corpus

DAVID HORNER & **PETER STRUTT**, UNIVERSITY OF LONDON BRITISH INSTITUTE IN PARIS

Introduction

This article reports on research to analyse domain-specific lexical categories using data derived from the Cambridge Learner Corpus (CLC). The CLC is a computerised database of contemporary written learner English, which currently stands at 17 million words and is being developed in collaboration with Cambridge University Press. The data comes from written BEC responses. BEC (Business English Certificate) is a suite of three business English examinations at levels B1 (BEC Preliminary), B2 (BEC Vantage) and C1 (BEC Higher) of the Common European Framework. The Sorted BEC Wordlist contains 9,437 headwords sorted alphabetically and by frequency of occurrence in the written scripts of candidates on the three BEC level examinations. No separation according to level is yet available although raw wordlists at each level will soon be produced.

Methodology

Firstly we identified four theoretical categories as a means of analysing lexis from the business domain. These categories were then tested against data derived from the sorted BEC Wordlist, firstly according to the judgements of the authors; then using the judgements of 28 learners of English working in a business environment, and six native speakers with no business experience.

We then compared these findings with data available in native speaker specialist corpora. Of particular interest were Nelson's (2000)[1] 1,023,000 running word list of business English derived from the British National Corpus and the Academic Word List (Coxhead 1998), made up of a balance of science, arts, commerce and law texts, and totalling 3,500,000 running words. The Academic Word List was developed at the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. The list contains 570 word families that were selected according to their frequency of use in academic texts. The list does not include words that are in the most frequent 2000 words of English. The AWL was primarily made so that it could be used by teachers as part of a programme preparing learners for tertiary level study or used by students working alone to learn the words most needed to study at tertiary institutions.

Developing lexical categories

Specialised or technical vocabulary is recognisably specific to a particular field, topic or discipline. The degree of specialisation depends on how restricted a word is to a particular area. It is possible to identify four main categories from most restricted to least restricted:

Category 1: the word rarely if ever occurs outside this domain. (most restricted)

For example:

Law	*obiter, mens rea, intestate*
Business	*dead cat bounce, bill of lading, cost-benefit analysis*
Computing	*pixel, serial port, run-time error*

Category 2: the word is used both inside and outside this field, but with a radical change in meaning.

For example:

Law	*consideration, execute (=perform a contract), cite (=appear)*
Business	*bull (on the stock market), interest (as in rate of interest), chair (person)*
Computing	*shell, protocol, field*

Category 3: the word is used both inside and outside this field; the specialised meaning can be inferred through its meaning outside the field.

For example:

Law	*accused (noun), reconstruction (of a crime), suspended (sentence)*
Business	*option (to buy or sell shares), drain (on resources), margin (profit margin)*
Computing	*cut and paste, bookmark, menu*

Category 4: the word is more common in the specialised field than elsewhere. (least restricted)

In this category there is little or no specialisation of meaning though someone knowledgeable in the field would have a more in-depth idea of its meaning. For example:

Law	*judge, trespass, act*
Business	*bank, discount, credit*
Computing	*icon, program, file*

[1] See, in particular, http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html; http://www.itri.brighton.ac.uk/-Adam.Kilgarriff/bnc-readme.html#bib; http://kielikanava.com/html; the Wolverhampton Business Corpus on http://www.elda.fr/cata/text/W0028.html; and http://users.utu.fi/micnel/business English lexis site.htm.

Although this might seem straightforward, when it comes to putting it into practice, a number of difficulties arise which we will discuss under each category.

Highly specialised as it is, one would expect category 1 to be, along with category 2, the smallest category. However, although many of the lexical items have domain restricted occurrence, the business world has become such a major part of our daily lives that few would not be recognised by non-specialists (e.g. *incoterms*), even if people would probably have only hazy ideas of the meaning of terms like *work-flow* and *outsource*. This is borne out by Nelson (op. cit.) whose 1 million word corpus of business English showed significant similarities, in terms of frequency, with general English.

Category 2 probably represents the most difficult for learners to cope with, since it contains apparently familiar words with highly domain-specific meanings, including *market* (not the one where you buy fruit and vegetables), *due* (owed), *brief* (not an adjective, but what one is instructed to do), *factor* (a means of obtaining payment), *asset* (a set of items on the balance sheet) and *goodwill* (an accounting value).

Categories 3 and 4 were expected to be the largest categories when applied to the Sorted BEC Wordlist. However, it is not always easy to draw the line between category 3 and categories 1 (at what point does an item begin to be obscure because of its domain specificity?) and 4 (at what point does a word become such a part of everyday usage that it loses its domain-specific validity?).

Hence one can wonder about the inclusion of the terms related to *market* and *advertisement*, which are so much part of the everyday lives of ordinary people that they might be said to have lost their domain-specificity. Other terms, like *deadline* or *regulation*, despite being in everyday use, are probably much more common in business-related environments. However, it is at this level that one really feels the need both for supporting evidence from other corpora, and for the possibility of looking at how the words are used in the Sorted BEC Wordlist, since collocational relationships probably play a major part in limiting domain-specificity. *Distribution* and *network* are in fairly common use, for instance, but it may be that they are used together in the unsorted learner corpus as *distribution network*. Similarly for *delivery* and *deadline*.

Category 4 presents the difficulty of opting for whether a word is in such regular everyday use as to have lost its domain-specificity. This is probably now the case with computer-related terms, but might not have been a decade ago.

There is the additional issue of polysemic words whose more common meaning would put them into this category, whereas their more restricted meaning would put them in category 1 or 2. Examples here would include *bond*, *share* and *bear*. Again, this is a factor signalled by Nelson (op. cit.), who found that:

- within a business setting the meaning potential of words – limited to a restricted number of semantic categories (people, companies, institutions, money, business events, places of business and lexis associated with technology) – was found to be reduced.
- certain key words are significantly more frequent within a business context than would be expected from their general frequency.

Both of these findings are significant when attempting to assign words to lexical categories. Before the findings are presented the final stage in the analysis should be described.

Simplifying the headword list

Because of the considerable difficulties related to assigning items to the four categories, it was decided to simplify the task facing our native speaker and non-native speaker business specialist informants. Firstly the list of 9,437 BEC headwords that had been provided by Cambridge ESOL was reduced to a shorter list of 632 items which was done in a number of stages:

- eliminating all the non-words and misspellings, including many words of probably foreign origin such as *AB* or *AG*;
- stripping away words with only one occurrence and irrelevant to business life, including many people and place names (*Briggs, Tyne, China*);
- eliminating all the high frequency grammatical words; deleting words like *desk, headlines, jet, request, communicate, reply* which do not belong to any specific domain;
- maintaining only a root word, unless the different morphological variants reflect business-important differences – thus, *commerce* and *commercialisation* were retained, but *commercialised* deleted;
- discounting adjectives and adverbs because, in isolation they are not marked for business use, although, obviously, a word such as *substantial* in the lexical unit *substantial growth* would be.

There is, however, an issue concerning adjectives and adverbs: although the vast majority of items are indeed non-domain specific, one feels that in many cases their relative frequency compared with a general English corpus is greater. This is almost certainly the case with compound adjectives, for instance, like *company-wide, day-to-day* or *well-established*. It is also probably true with many items in more common use but which may have strong collocational links with verbs and nouns in frequent use in the business domain – *considerable/dramatic/drastic* rise/fall; yours *faithfully/sincerely/truly*. Indeed, Nelson (op. cit.) notes that lexical items within the business domain tend to fall within distinct semantic categories. The adjectives and adverbs in the Sorted BEC Wordlist can often be quite readily associated with his categories, but others also spring to mind: talking about figures and trends – *accurate, adequate, approximately*; describing companies, institutions and people – *active, conservative, corporate*; letter language – *above-mentioned, afraid, dear*; location – *abroad, African, company-wide*; describing behaviour – *acceptable, advisable, appropriately*.

The list established after discussion was then submitted to 28 learners of English (NNS) with extensive business experience in a multinational context[2], and six native speakers with no business experience[3] (NS). These informants were then offered the alphabetically sorted short list and asked to identify:

- those words which were highly specific to a business context;
- those words which, although in more frequent use in business contexts are also well-known to the general public;
- and they were asked to leave unmarked any words which they considered to be simply general English.

The results are reported below based on the students' responses to the first 126 items (approximately 20% of the list).

Results

The analysis of the top fifth of items on the short BEC word list are revealing. It is striking that not one single item was selected by all 28 NNSs, and only five (four business-specific and one non-specific) by more than 22, and seven (three business core and four non-specific) by more than 17. There is thus some degree of unanimity on only 12 items (9.5% of the short list). On the other hand, opinions were split for the other words. This confirms that the NNSs were having the same difficulty as the experts in clearly identifying core items. The following table shows the number of items selected by all, some, or none of the NNS:

Table 1: Number of items selected by the NNS group

No. of items selected for each range in each category (N=126)		Range of NNS who chose them (N=28)
Business specific	Non-specific	
0	0	28
4	1	23–27
3	4	18–22
11	15	13–17
39	58	8–12
32	29	3–7
0	15	1–2
37	4	0

The most frequently identified core items were: *audit, auditor, bookkeeping* and *ceo* (23–27 respondents) and *accountant, banking* and *capitalisation* (18–22 respondents). There is significant overlap here with the six NSs, who all selected *audit, auditor, capitalisation* and *ceo*. Of the other two selected by all six NSs, however – *appraisal* was surprisingly selected by only 3–7 of the NNSs, and *automate* by none.

The six NS respondents showed more agreement: 39 items were placed in the same category by all of them – 33 in non-specific (category 2) and 6 in business specific (category 1). On the other hand, this means there was still disagreement about the majority of items, with a 50% split on 42 items, while 46 of the remaining 97 items were all chosen at least once.

Table 2: Number of items selected by the NS group

No. of items selected per NS (N=126)		Number of NS who chose them (N=6)
Business specific	Non-specific	
6	33	6
0	14	5
0	4	4
23	19	3
0	18	2
0	28	1
97	10	0

It would appear therefore that both our NNS and NS respondents were having similar difficulties to us in assigning categories with confidence.

Conclusion

This research has provided informative insights into developing and using lexical categories and into the venture of classifying domain specific lexis according to meaning-based categories. The fuzziness of the categories proposed is clear. Future studies by Cambridge ESOL will use corpus data as evidence of domain-specific lexis in use.

References and further reading

Coxhead, A (1998): *An Academic Word List*, Occasional Publication 18, LALS, Victoria University of Wellington, New Zealand.

McCarthy, M (1991): *Discourse Analysis for Language Teachers*, Cambridge: Cambridge University Press.

Nation, P (2001): *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.

Nelson, M (2000): *A Corpus-Based Study of Business English and Business English Teaching Materials*, Unpublished PhD Thesis, Manchester, University of Manchester.

Sutarsyah C, Nation, P and Kennedy, G (1994): How useful is EAP vocabulary for ESP? A corpus-based study, *RELC Journal* 25, 34–50.

[2] 28 intermediate (BEC Vantage) and advanced (BEC Higher) students of business English at the British Institute in Paris.

[3] Six second- and third-year students of French at the British Institute in Paris.

IELTS Writing: revising assessment criteria and scales (concluding Phase 2)

STUART D SHAW, RESEARCH AND VALIDATION GROUP

Introduction

Phase 1 of the IELTS Writing Assessment Revision Project – *Consultation, Initial Planning and Design* – was reported in *Research Notes 9*. Progress on the second phase of the project – the *Development* Phase – was described in issue 10. This article presents a résumé of outcomes for both phases culminating in a discussion of a number of Phase 2 issues relating to underlength or illegible scripts, memorisation and task rubrics.

Phase 1 – Consultation, Initial Planning and Design

Phase 1 of the project involved consultation with a range of stakeholders and was completed in December 2001. Initial discussion within the Revision Working Group was informed by a review of studies relating to IELTS Writing, and also by a comprehensive survey of the literature on holistic and analytic approaches to writing assessment. The next step was to explore current practice among IELTS Writing assessors, in order to gauge their attitudes towards their respective assessment practice and to highlight theoretical and practical factors which would help shape the redevelopment of the writing assessment criteria and scales.

The consultation phase began with a series of semi-structured interviews with groups of IELTS Academic and General Training Writing assessors in the UK and Australia. These interactions led to the construction of a survey questionnaire which was sent out to a sample of several hundred IELTS assessors based at a range of test centres worldwide. The purpose of the interviews and questionnaires was to elicit from assessors individual approaches and attitudes to the assessment of IELTS writing tests, especially in relation to differing domains (Academic and General Training) and differing task genres (Task 1 and Task 2). Protocol analyses are capable of revealing rich insights on the part of assessors which can be instrumental in helping to develop assessment criteria and scales that are valid, reliable and practical.

The questionnaire consisted of sections exploring assessors' approaches and attitudes to:

- rating the different task types for Task 1 and Task 2;
- using Global and Profile scales;
- interpreting the assessment criteria and band descriptors.

Phase 2 – Development

The *Development* Phase of the project was completed in May 2003. The combined use of quantitative methodologies (application of draft criteria and scales to sample language performance) and qualitative methodologies (insightful and intuitive judgements derived from 'expert' participants) informed the re-construction of assessment criteria and scales for the IELTS Writing Test.

Four key revision areas were identified during the *Development* Phase:

Assessment approach

The benefits of analytical assessment in relation to the IELTS examination – enhanced reliability through increased observations; wide range of writing performances; greater discrimination across wider range of assessment bands (9 Bands); provision of a greater control over what informs the impressions of raters; removal of the tendency to assess impressionistically; active discouragement of norm-referencing and the provision of research data/information – suggested that analytic assessment outweighed any advantages offered by a global approach to assessment.

Assessment criteria

Enough similarity in the two writing tasks exists across the Academic and General Training Modules to warrant the use of the one set of assessment criteria for each rather than developing separate criteria. Consequently, a revised set of criteria was developed for Task 1 in both Academic and General Training Modules and a separate set developed for Task 2 in both modules.

The five revised criteria for both Modules and both Tasks are: *Task Achievement* (Task 1)/*Task Response* (Task 2), *Coherence and Cohesion* (Task 1 and 2), *Lexical Resource* (Task 1 and Task 2) and *Grammatical Range and Accuracy* (Task 1 and Task 2).

Rating scale descriptors

The band descriptors evolved through a succession of iterative drafts and fine tunings, the final form being an amalgamation of expert contributions which is widely regarded as a rater-friendly instrument.

Examiner training

It will be necessary through the implementation of new training systems to re-train all writing examiners subsequent to any revision. In addition, it will be a requirement for all examiners to re-certificate. Any changes to the Writing Assessment Guidelines (WAG) generated by the project will be included in a revised document before re-certification. Documents affected by the revision (examiner training packs, certification sets, revised WAG, revised specimen materials and new item writer specifications) will need to be prepared in time for training and certification to take place prior to the launch.

Outstanding Phase 2 issues

Conclusion of the second phase of the project comprised consideration of a range of script and task-related issues and outlining the Validation issues which will form Phase 3 of this project. The script and task-related issues are underlength scripts, illegible scripts, memorised scripts and instructions in the task rubric, each of which will be discussed further below.

Underlength scripts

The issue of underlength scripts, that is, scripts which are either incomplete, unfinished or unduly short becomes especially pertinent in cases where task fulfilment is an integral part of the scoring rubric. In the case of incomplete or unfinished scripts raters are clearly faced with a dilemma: if the test taker produces a promising beginning to a script but fails to complete it, the rater is left with a decision. Does the rater award a score based on the linguistic quality of what has been written, assuming of course that the writer could have made an appropriate conclusion given more time, or does the rater adhere rigorously and rigidly to the wording of the task rubric and rate the script on the basis of what is present, especially when the rubric provides explicit instructions regarding response length.

In the case of a high-stakes test like IELTS, it is important to consider both the purpose of the assessment and the impact that any decisions taken will have on test takers. IELTS provides, amongst other things, an indication of a test taker's ability to produce writing in an academic context. In this sense, strict rating criteria are justified and, consequently, underlength scripts should be penalised.

Moreover, it is essential that raters are made aware of and anticipate problems associated with brief responses. As a consequence, IELTS raters should be given advice on how to deal with extremely brief responses, or responses in which the writer has demonstrated an understanding of the salient features of a task but was unable to complete the task in the allotted time.

The most appropriate and equitable way of penalising underlength scripts is to employ a range of empirically-informed penalties for scripts of varying length. A 3-band penalty system under *Task Achievement*/*Task Response* will be imposed. Amongst the recommendations made by the Revision Working Group, the following have been proposed:

- underlength responses to all IELTS Writing Tasks should continue to be penalised using the revised scales (an underlength response can be thought of as less than 150 words for Task 1 and less than 250 words for Task 2);
- a sliding scale system should be imposed where a fixed penalty is applied to a response comprising a word length falling within a specific range;
- answers to Task 1 and 2 (for both Writing Modules) which are two lines or less will be automatically scored as Global Band 1;
- underlength penalties should be communicated to writing examiners by placing an explicit statement of quantification outside the band descriptors i.e. locate a response within a band descriptor and then invoke a penalty from a list of specific penalties for varying degrees of underlengthness.

Memorised scripts

Determining task fulfilment is especially difficult in the case of memorised scripts. It is clear that a memorised script does not provide an accurate writing sample of a test taker's ability as there is no way of knowing either the author or the source. It is, therefore, crucial that certain steps be taken to avoid the possibility of memorised scripts being proffered or accepted for rating. Some adherence to the task in order for a script to be deemed acceptable is undoubtedly one way of minimising memorisation and in general, the more specifically a writing task is tailored towards a given situation, the more important the notion of task fulfilment becomes.

Since September 2003 all Writing Tasks used for the IELTS examination form part of a discrete test and, therefore, the issue of whole script memorisation is now substantially less of an issue than it was. Totally memorised scripts – where evidence of plagiarism can be obtained – will continue to receive Band 0. The issue of memorised phrases i.e. the regurgitation of prefabricated lexical chunks, remains an obdurate issue and continues to be dealt with by the current band descriptors. Now any cases of memorised responses to a topic are penalised under *Task Achievement* or *Task Response* for irrelevance.

Illegible scripts

Recent studies by Brown (2000) and Shaw (2003) have investigated differences between handwritten and word-processed versions of the same essay. Brown looked at IELTS Task 2 essays and the effects of handwriting on legibility and assessment and deduced that legibility has a marginal but significant impact on scores. Moreover, the size of the effect is relative to the quality of handwriting and neatness of presentation. Contrary to her hypotheses, the handwritten versions of the same script were assessed higher than the word-processed versions : the worse the handwriting – the higher the comparative assessment. Shaw, using FCE scripts, arrived at a similar finding (see *Research Notes 11*).

In second language writing assessment, it may be the case that greater stress is put on certain linguistic features such as grammatical accuracy and range, lexical resource and syntactical structures. Unlike first language assessment, 'mechanical' aspects of writing such as mastery of orthographic and iconic conventions and handwriting neatness may not be particularly significant assessment foci. There exists in first language assessment, on the other hand, a marked central focus on mechanical aspects of writing. Poor legibility might well serve to distract from mechanical errors. Anecdotal evidence suggests that second language raters may even attempt to compensate for poor legibility by more careful and attentive reading in the hope of avoiding discriminating against candidates because of poor script legibility.

It is rare for a script to be totally illegible. Where they exist, such scripts are impossible to read or understand and, therefore, cannot be rated. It is normal practice to award completely illegible scripts a Band 1. However, it is proposed that illegible scripts should be read by three examiners. If all three examiners consider the script totally illegible then a Task Band Score of 1 should be made. Partially legible scripts can only be rated against retrievable language and self-penalise under *Coherence and Cohesion*. A second examiner should attempt to rate a partially legible script before awarding a final band. Unreadable sections will not be discounted under word count.

Instructions in the task rubric

The test rubric comprises certain characteristics or facets that specify how the test taker is expected to undertake the test. Bachman (1990:118) suggests that these characteristics include the *test organisation, time allocation* and *instructions*. It is the last characteristic which provided the focus for Phase 2 discussions on the current task rubrics.

According to Bachman (1995:124):

> *In general, the more complex the task required and the less familiar it is to the test takers, the greater the burden carried by the* ***instructions*** *to specify clearly what is expected of the test taker.*

One source of test taker anxiety, according to Madsen (1982), are unclear instructions. Providing clear instructions for test takers is a crucially important aspect in designing and implementing a valid test. Bachman and Palmer (1996) offer three essential guidelines for instructions. They should be:

- simple enough for test takers to understand;
- short enough so as not to take up too much of the test administration time; and
- sufficiently detailed for test takers to know exactly what is expected of them.

For a direct test of writing, such as the IELTS Writing Test, minimum requirements for task instructions should include:

- some indication of the purpose for writing;
- some indication of how the writing will be assessed; and
- some further indication of the response length.

The instruction on length should be in the form of either a minimum word count, a structural unit (such as sentences or paragraphs) or, as has been suggested by Carson (2000), page units (e.g. one or two pages). In the case of IELTS Writing there is currently a minimum word count – 150 words for Task 1 and 250 words for Task 2 (on both Writing Modules).

It is generally believed that what in fact constitutes task fulfilment lies in the purpose of the assessment and the type of rating employed. For example, with assessments whose main purpose is obtaining a rateable sample of writing that will exhibit control of syntax and vocabulary, the extent to which writers follow task instructions will be less important than for assessments whose primary function is to assess writers' ability to successfully communicate in writing, such as those typical of academic settings.

Some minor changes to the current rubrics for *Academic Writing* Tasks 1 and 2 are currently under consideration but will not be implemented immediately. Slight textual amendments to the instructions have already been agreed to ensure that candidates are fully aware of task requirements. Since January 2004 there has been a warning on the cover sheet of the Writing paper (under *Information for candidates*) informing candidates that:

- underlength scripts will be penalised;
- they must attempt both tasks;
- they write at least 150/250 words for Task 1 and Task 2 respectively.

The new script for invigilators will also refer to the importance of candidates writing the requisite number of words or they will be penalised as explained above.

Phase 3 – Validation

A validation programme in support of Phase 3 has been organised around the major headings which constitute four essential examination qualities: validity, reliability, impact and practicality (VRIP) with subheadings adapted from the VRIP checklists used, for example, in the CPE Revision Project and in establishing CELS. Successful validation of the revised rating scale cannot be accomplished without due consideration being given to all four VRIP components. Results of these validation trials will be described in *Research Notes 16*.

References and further reading

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

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

Brown, A (2000): *Legibility and the rating of second language writing: an investigation of the rating of handwritten and word-processed IELTS Task Two essays*, IELTS Research Reports Volume 4.

Carson, J G (1999): Reading and writing for academic purposes, in Pally, M (ed): *Sustained content teaching in academic ESL/EFL*, Oxford: Houghton Mifflin, 19–34.

Madsen, H S (1982): Determining the debilitative impact of test anxiety, *Language Learning* 32,133–43.

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, Cambridge: UCLES.

An IELTS Impact Study: implementation and some early findings

ROGER HAWKEY, CONSULTANT, CAMBRIDGE ESOL

Introduction

Impact studies in the field of education normally focus on the effects of *interventions*, including teaching programmes and tests, on people participating in them in various ways. Such studies measure, describe and evaluate not only *outcomes*, for example, test results or subsequent performance on the criteria the test is measuring, but also *processes*, for example the learning and teaching on programmes preparing candidates for the test, and *perceptions*, for example attitudes engendered by the test. Weiss (1998) sees *impact* as referring to the effects of a programme (or test) on the larger community, noting that impact may be planned or unplanned; positive or negative; achieved immediately or only after some time; and sustainable or unsustainable.

This article, the third on the IELTS Impact Study (IIS) in *Research Notes*, will describe the implementation of the study and report some of its early findings.

The IELTS Impact Study: Phase 1

In *Research Notes 2*, Nick Saville described the IELTS Impact Study from its inception in 1995, when, he noted, “it was agreed that procedures would be developed to monitor the impact of the test and to contribute to the next revision cycle”. Saville then explained the rationale for IELTS impact studies:

> *In order to understand the test impact better and to conduct effective surveys to monitor it, it was decided that a range of standardised instruments and procedures should be developed to focus on the following aspects of the test:*
>
> - *the content and nature of classroom activity in IELTS-related classes*
> - *the content and nature of IELTS teaching materials, including textbooks* (see also Saville and Hawkey 2003)
> - *the views and attitudes of user groups towards IELTS*
> - *the IELTS test-taking population and the use of results* (2001:5).

The IIS is thus broad in scope, covering the impact of the IELTS on a range of stakeholders, in the classroom and beyond. Its major focus is on the washback of the test on language teaching and learning, taking account, as suggested by Milanovic and Saville (1996:2), of the “complex interactions between the factors which make up the teaching/learning context (including the individual learner, the teacher, the classroom environment, the choice and use of materials etc)…”.

The IIS has now completed the third of its three phases, as summarised by Saville in Figure 1.

Figure 1: The three phases of the IELTS Impact Study

Phase 1	Identification of areas to be targeted and the development of instrumentation to collect information which allows impact to be measured
Phase 2	Validation of the instruments prior to full-scale implementation
Phase 3	Implementation of the instruments as part of a major survey

The initial research for Phase 1 of the study was undertaken on commission from Cambridge ESOL by Professor Charles Alderson at Lancaster University (see, for example, reports to Cambridge ESOL by Alderson and his team in: Alderson and Banerjee 1996; Banerjee 1996; Bonkowski 1996; Herington 1996; Horak 1996; Winetroube 1997; Yue 1996).

IIS: Phase 2

As a consultant invited to work on the implementation of the IIS, I described, in *Research Notes 6*, the main developments of Phase 2. This saw extensive analyses and pre-testing of the draft data collection instruments by the Validation Group, with consultancy support from, among others, Professor Lyle Bachman, Dr Jim Purpura, Professor Antony Kunnan, and myself. In the process of Phase 2, the original thirteen data collection instruments were rationalised into five:

- a modular student questionnaire on pre- and post-IELTS candidate language learning background, aims and strategies; test-preparation programmes, and IELTS attitudes and experience;
- a language teacher questionnaire, covering teacher background, views on IELTS, experience of and ideas on IELTS-preparation programmes;
- an instrument for the evaluation of IELTS-related textbooks and other materials (see Saville and Hawkey 2003);
- a classroom observation instrument for the analysis of IELTS-preparation lessons;
- a *pro forma* for receiving institute IELTS administrators on their IELTS experiences and attitudes.

In May 2001, a pre-survey questionnaire was sent to a world-wide sample of over 300 University, British Council, IDP Education Australia and other test centres world-wide. Responses to this pre-survey were obtained from 41 countries and gave information on: the language tests for which each centre runs courses; numbers, durations and dates of such courses; numbers and nationalities of students; textbooks and other materials used.

These data were used to help select the IELTS centres for the main data-collecting Phase 3 of the IIS.

IIS: Phase 3 approaches

The research methodology of the IIS has always combined quantitative and qualitative data collection and analysis, with the balance towards the qualitative end of the continuum.

The IIS employs *survey research* approaches, in its use of questionnaires, structured and in-depth interviews and observation. Survey research is not, of course, confined to large-sample studies and simple statistical analyses. Its approaches suit studies like the IIS, seeking probabilistic and interactive, not necessarily deterministic, relationships between individual and group characteristics such as language background and proficiency, attitudes, motivation; language learning approaches, strategies and styles of language teaching (see Baker 1997:35).

The IIS attempts to minimise problems sometimes associated with survey research. These (Baker op. cit.) can include a lack of clear aims, implicit rather than explicit theoretical input, inability to establish causal relationships, inadequate sampling, instruments containing invalid items, lack of triangulation through other data collection methods, and interviewer or researcher effects. The IIS has adhered to the objectives summarised by Saville above, states its various research hypotheses, uses validated data collection instruments, triangulates data (e.g. teacher questionnaire responses and teacher interviews, student and teacher questionnaires with classroom observation analyses), acknowledges the limitations of its samples and the context- bound nature of its findings, and, where appropriate, refers data to further statistical analyses.
All these processes are described in the full IIS report. But the study also seeks to fulfil its potential “to examine the interplay of variables and issues to a greater extent than quantitative research typically sets out to do” (Duff 2002:6) “The latter”, Duff adds, “conversely attempts to control as many extraneous variables as possible”.

The summary below focuses on early IIS findings in selected areas only, that is IELTS impacts on preparation courses, test module difficulty and perceptions of and pressures from the test.

The IIS participants

From the pre-survey data, a case-study sample of around 30 centres was selected, closely representative of the IELTS nationality population, and IELTS candidates and teachers at these centres contacted with the invitation to respond to the IIS instruments. To date, we have received responses from 572 IELTS candidates, from 83 teachers completing the teacher questionnaire, and from 43 teachers completing the instrument for the analysis of textbook materials; 12 IELTS-preparation classes have been recorded for analysis. Triangulation of both the closed- and open-ended data from the questionnaires has been attempted through stakeholder (student, teacher, administrator) interviews and focus groups, and through classroom observation at selected case-study centres, involving 120 students, 21 teachers and 15 receiving institution administrators.

Early data analyses from the IIS questionnaire, interview and classroom observation data indicate useful findings on the backgrounds, aspirations, attitudes and perceptions of those affected by IELTS. Some of the key early findings are illustrated here, to be confirmed and supported with appropriate statistical detail in the full report to Cambridge ESOL.

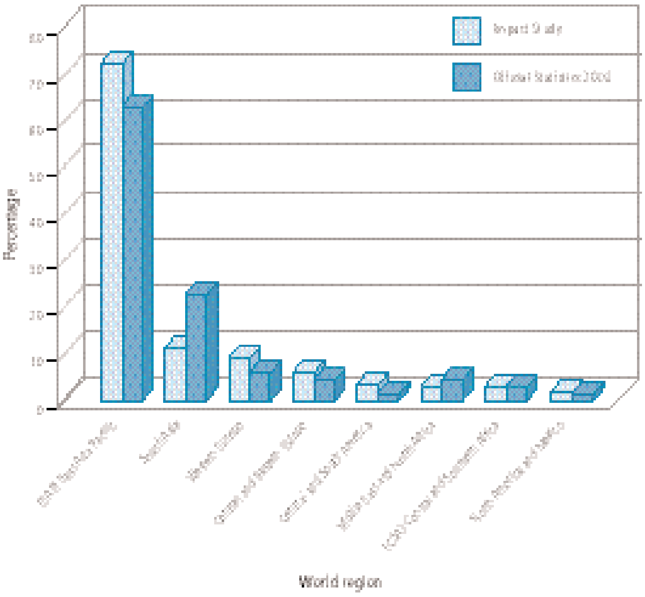
The base characteristics of the IIS student population are summarised in Table 1:

Table 1: IIS candidate population: some baseline data

Gender %	Female	55
	Male	45
IELTS Status %	Pre-IELTS	64
	Post-IELTS	36
Module %	Academic Module	89
(of post IELTS	General Training Module	11
sub-population)		
Educational	Postgraduate	48
Level %	Undergraduate	46
	Pre-university	6
Fields %	Business	30
	Finance	14
	ITC	7
	Law	4
	Hotels, Tourism	2
	Health & Social Sciences	17
	Education	13
	Engineering	6
	Public Administration	3

The world regional proportions of the IIS candidate population is compared with the 2002 IELTS candidacy in Figure 2, indicating a fair match apart from some discrepancy in the East Asia Pacific, South Asia and Western Europe proportions.

Figure 2: Comparison between IIS population distribution (by world region) and actual IELTS candidacy figures for 2002



IELTS impacts on preparation courses

The analyses of the teacher questionnaires (triangulated with findings from analyses of the video-recordings of IELTS preparation classes) are producing interesting findings on the influences of the test on language teaching and learning approaches and materials. IIS student participants attended the following types of language preparation courses pre-IELTS.

Table 2: Pre-IELTS course types and main activities

Preparation course types (n = 233)	
IELTS-specific	108
EAP/Study skills	86
General English	39

Ninety per cent of the IIS participant teachers agreed that the test influences the content of their lessons, 63% that it also influences their methodology, leading to preparation programmes that are more focused, mainly test-oriented, aimed at developing relevant communicative micro-skills, encouraging discussion and brainstorming, and using authentic texts, including a wide range of multi-media target language materials from beyond the textbooks. Of the specific references made by teachers in response to an open-ended item on changes in their methodology for IELTS-preparation courses, references to practising candidates in *time-constrained* activities were significantly frequent. According to responses to the student questionnaire, the main test-specific activities on preparation courses are practice tests and the consideration of past papers, followed by work on exam techniques and discussion of test format.

Table 3 summarises the responses of 431 IELTS candidates to open-ended student questionnaire items on their perceptions of success on their IELTS preparation courses.

Table 3: Candidate views on their success on IELTS preparation courses

Do you think you were/are successful on the preparation course(s)?	
Yes	184
No	188
Unsure	39
Other	20

The main criteria for success mentioned by those responding *positively* to this item were: perceived improvement in English proficiency level (26 responses) or skills (14 responses) and increased familiarity with the test (33 positive responses). This result suggests an interesting balance in students' perceptions of success between improvement in their target language proficiency, and increased test familiarity. Of the *negative* responses, 23 candidates felt they had not had enough time to prepare for the test, 18 that they had not worked hard enough and 15 that their courses had not provided enough IELTS practice or preparation.

Teachers completing the IIS instrument for the evaluation of textbook materials add interesting insights here, noting that the IELTS-related textbooks they use cover, in particular, micro-skills such as: identifying main points, identifying overall meaning, predicting information, retrieving and stating factual information, planning and organising information, and distinguishing fact from opinion. While IELTS appears to encourage communicative activities across the four skills in line with the test tasks, the opportunities for learners to communicate on their own behalf during their preparation classes seem, according to the IELTS-preparation lessons recorded and analysed, to vary very considerably.

To a set of questions designed to cast further light on the nature and strength of the impact of IELTS on preparation courses, the teachers responded as in Table 4 here.

Table 4: Teacher responses to items on IELTS and non-IELTS preparation

If an IELTS score had not been a requirement would you have prepared your students for their future studies in the same way?	
Yes	32%
No	68%
Would your IELTS preparation course be a good way to learn English for someone going to University but who is not going to take IELTS?	
Yes	63%
No	37%
Would the IELTS preparation course be useful for someone who is not going to University?	
Yes	54%
No	46%

These responses, along with other IIS evidence, suggest that IELTS does indeed impact significantly on preparation courses, but that the resultant programmes are considered reasonably appropriate for students preparing to enter university and for those who are not.

Test module difficulty

Table 5 indicates that both candidates and IELTS preparation teachers have similar perceptions on the relative *difficulties* of the IELTS skills modules.

Table 5: IIS student and teacher perceptions of IELTS module difficulty

Most difficult IELTS Module? (%)		
	Students	Teachers
Reading	49	45
Writing	24	26
Listening	18	20
Speaking	9	9

This variation across the perceived difficulty of the skill modules is reflected to some extent in the preparation courses. Table 6 is evidence that reading activities are seen by the candidates as occupying an average of 26% of course time, compared with almost equal proportions for listening, speaking and writing.

Table 6: Candidate perceptions of preparation course activities

Preparation course activities (%)	
Reading	26
Writing	16
Listening	19
Speaking	17
Vocabulary	7
Grammar	7
Other	1

These samples of early IIS survey data on the four macro-skills, combined with related interview and classroom observation analyses, will be useful for validation work on whether IELTS reading comprehension is and/or should be harder than the other skills tests. Detailed discussions on the skills modules at the interviews and focus groups with candidates and teachers, included questions on whether the reading and writing topics are too humanities oriented, and not specialised enough; whether some skill module tasks are too formulaic and predictable, and, above all, the extent to which the tests reflect appropriate processes and target institution activities. One radical suggestion made at a focus group meeting was that IELTS should be skill-modular (like the Certificates in English Language Skills exam (CELS)), with skill module(s) to be selected as required by candidates, who would be able to accumulate elements of the test score over a period of time. This, it was claimed, would allow people who fail marginally on one element to focus on the relevant area before retaking the exam three months later.

Test perceptions and pressures

IIS participants who had already taken IELTS were asked whether they thought IELTS a fair way to test their proficiency in English. Seventy per cent (of the 190 concerned) responded yes. Of the reasons given for the negative responses, the most common, in order, were: opposition to *all* tests; pressure, especially time constraints, and the influence of test topics. The absence of a grammar test, writing and speaking test rating and the brevity of the speaking test were less frequent but nonetheless relevant areas of concern. Candidates clearly feel the pressure of taking a high-stakes exam such as the IELTS, around one-third claiming to feel "very worried" before the test. Though 53% of the IIS *teachers* were concerned that IELTS caused candidates some stress, 94% also claimed that the test provided a positive motivation for candidates.

Fifty-four per cent of the IIS post-IELTS participants did not feel that they had performed to the best of their ability on the test. Table 7 summarises aspects of the test that candidates perceived as most affecting their performance in it:

Table 7: Candidates' post-IELTS perceptions of what affected their performance most

Factors affecting IELTS candidate performance (%)	
Time pressure	40
Unfamiliarity of topics	21
Difficulty of questions	15
Fear of tests	13
Difficulty of language	9

IIS figures on IELTS band scores show interesting relationships between the scores required by target institutions, candidate expectations and actual results. The mode bands of the relevant sub-groups of the IIS student populations were as follows: band 6 for the scores already achieved, band 6.5 for the bands stated as required, and band 6 as the band expected. These data will be analysed further along with qualitative information on target institutional and departmental claimed and actual IELTS entry score cut-off points, and quantitative data on IELTS candidate average score tendencies over time.

Conclusion

In addition to the impact areas selected for attention here, the full IIS report will provide insights at the IELTS validity:reliability interface. The findings are not, of course, to be pre-empted, but partial analyses of data from the candidate, teacher and administrator suggest perceptions that IELTS:

- is considered a reasonable direct communicative performance test, appropriate for use with candidates for under-graduate and post-graduate studies, and for those seeking an English qualification for professional purposes;
- has content most of which seems relevant to target communicative activities, e.g. authentic texts, a range of

micro-skills, but IELTS writing (and reading) tasks are considered 'too general' by some and/or may not be relevant to all candidates.

From the analyses of the qualitative and quantitative data collected, hypotheses will be developed on many aspects of IELTS impact. Findings and recommendations that are felt to need further inquiry will be compared with related IELTS research (e.g. Green 2003, Read and Hayes 2000) or receive it in a possible Phase 4 of the impact study.

References and further reading

Alderson, J C and Banerjee, J (1996): *How might Impact study instruments be validated?* A paper commissioned by UCLES as part of the IELTS Impact Study.

Baker, C (1997): Survey Methods in Researching Language and Education, in Hornberger and Corson (eds): *Encyclopedia of Language and Education Volume 8: Research Methods in Language and Education*, Dordrecht: Kluwer Academic Publishers.

Banerjee, J V (1996): *The Design of the Classroom Observation Instruments*, internal report, Cambridge: UCLES.

Bonkowski, F (1996): *Instrument for the Assessment of Teaching Materials*, unpublished MA assignment, Lancaster University.

Duff, P A (2002): *Beyond Generalizability: Contextualization, Complexity, and Credibility in Applied Linguistics Research*, paper delivered at the American Association of Applied Linguists/International Language Association Conference, 6 April 2002.

Green, A (2003): *Test Impact and EAP: a comparative study in backwash between IELTS preparation and university pre-sessional courses*, research for the Ph.D degree of the University of Surrey at Roehampton.

Herington, R (1996): *Test-taking strategies and second language proficiency: Is there a relationship?* Unpublished MA Dissertation, Lancaster University.

Horak, T (1996): *IELTS Impact Study Project*, unpublished MA assignment, Lancaster University.

Lazaraton, A (2001): *Qualitative research methods in language test development*, paper delivered at the ALTE Conference, Barcelona, July 2001.

Milanovic, M and Saville, N (1996): *Considering the impact of Cambridge EFL examinations*, internal report, Cambridge: UCLES.

Read, J and Hayes, B (2000): *The impact of IELTS on preparation for academic study in New Zealand*, report for the IELTS Research Committee.

Saville, N and Hawkey, R (2003): A study of the impact of the International English Language Testing System, with special reference to its washback on classroom materials, in Cheng, L, Watanabe, Y and Curtis, A (eds): *Concept and method in washback studies: the influence of language testing on teaching and learning*, Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Weiss, C (1998): *Evaluation*, New Jersey: Prentice Hall.

Winetroube, S (1997): *The design of the teachers' attitude questionnaire*, internal report, Cambridge: UCLES.

Yue, W (1997): *An investigation of textbook materials designed to prepare students for the IELTS test: a study of washback*, unpublished MA dissertation, Lancaster University.

The YLE Review: findings from a stakeholder survey

TRISH BURROW AND **JULIET WILSON**, EXAMINATIONS AND ASSESSMENT GROUP

Introduction

As part of Cambridge ESOL's ongoing commitment to improving the quality of its tests, we are currently conducting a Review of the Young Learners English Tests (YLE). We are following the model of test development and revision described in a previous article in *Research Notes 4* (Saville 2001). In this article we report on the consultation process, which is the first stage of any exam review or revision.

The YLE Review

The YLE tests were developed between 1993 and 1997. The test development team worked closely with staff from Homerton College (Cambridge University's teacher training college) to produce a test that took account of: current approaches to curriculum design and pedagogy for young learners; children's cognitive and first language development and the potential influence of test methods. The first administrations of the YLE tests were carried out in 1997 and this is the first review of the tests.

In May 2003, stakeholders were asked for their opinions on the YLE tests. We sent out two questionnaires, one to Local Secretaries in all YLE centres and one to all Senior Team Leaders. Local Secretaries were asked for feedback on aspects such as the content and format of each paper, administration procedures and the impact of the tests on classrooms around the world. We asked Senior Team Leaders to give feedback on aspects such as the appropriacy of task types and interlocutor frames used in the Speaking Tests, the quality of current training materials and assessment criteria, and the cultural appropriacy of the test materials.

Both Local Secretaries and Senior Team Leaders provided us with valuable feedback about all aspects of YLE. A total of 98 centres responded; this included all of our 20 largest centres.

Table 1: Mean score of responses from Local Secretaries

	Starters	Movers	Flyers	Average
Overall satisfaction with YLE Tests	4.44	4.50	4.44	4.46
Reading/Writing Test	4.43	4.46	4.26	4.38
Listening Test	4.43	4.46	4.44	4.45
Speaking Test	4.24	4.24	4.35	4.28
Administration procedures	4.47	4.48	4.41	4.45
How far do the tests meet learning goals?	4.36	4.28	4.13	4.26
How well do the tests reflect classroom practice?	4.06	3.93	3.89	3.96
How positive is the impact on classroom practice?	3.82	3.72	3.68	3.74

All Senior Team Leaders working in regions where the YLE Tests are taken responded. Senior Team Leaders whose geographical areas contain several countries further asked a selection of Team Leaders from these countries to join in the consultation process.

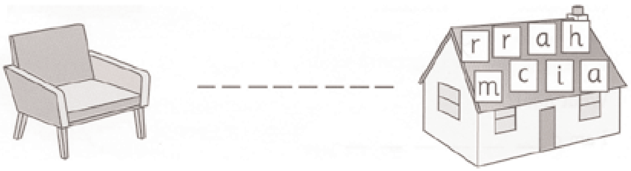
Satisfaction measures

Centres were asked to provide a satisfaction rating, ranging from '1 = not at all satisfied' to '5 = very satisfied' on a range of topics (questions 1–5). Three additional questions (numbers 6–8) related to the tests and their impact on the classroom, and for these questions the rating ranged from '1 = not at all' to '5 = very well'. Table 1 shows the mean score of all the responses received from Local Secretaries.

As can be seen, there is a high level of satisfaction with the Young Learners English Tests. Whilst further improvements can always be made, we were very pleased to hear that our key stakeholders have a very positive view of the tests. In order to ensure continued stakeholder satisfaction, we do not intend to make any major changes to the test. Many of the tasks may remain unchanged while others will feature minor modifications. There were, however, some areas highlighted by centres in the consultation process that we have decided to investigate further.

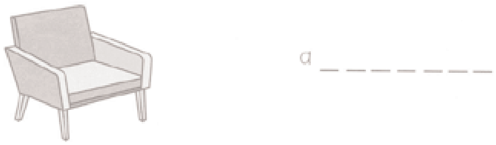
Listening and Reading/Writing

As part of the development phase of the Review, we are currently trialling new variants of some existing tasks. For example, in Part 3 of the Starters Reading & Writing paper spelling and written production of vocabulary is tested. In the existing task the children are required to write five words from a given lexical set. For each item, the children are supplied with a picture, the appropriate number of spaces and a set of jumbled letters. For example,



We wanted to research the effect of not providing the children with the jumbled sets of letters, as it seems that some children try to use all the letters in the anagram, but do so incorrectly. For example, for 'armchair', attempts at the word can include *charrima and *armachir.

In the revised variant of this task, the children are required to write the words and are supplied with a picture, the first letter of the word and the correct number of spaces in which to write the word. No jumbled letters are given. For example,



This and other revised tasks were trialled in several of our key markets between May and August 2003. In addition to this, centres participating in the trialling were sent a questionnaire and asked to comment on the format of the revised tasks. Cambridge ESOL sees this an important qualitative aspect of the trialling.

Speaking

As mentioned above, another part of our review process included consultation with the Senior Team Leaders who are responsible for the training and monitoring of Oral Examiners. We asked them about their overall satisfaction with the Speaking Tests and they also commented in more detail on individual tasks and assessment procedures. Table 2 shows their mean ratings of the speaking tests (on the same scales as the Local Secretaries' table above).

In general the feedback from Senior Team Leaders was positive, although a few areas for improvement were highlighted.

A selection of revised Speaking tasks were developed and subsequently trialled with young learners in July and August 2003. In particular we are looking at the storytelling task and ways of clarifying the construct and what the children are expected to achieve in this part (see Ball 2001 for a previous study).

Table 2: Mean score of responses from Senior Team Leaders

Question	Mean score
How satisfied are you overall with the YLE Speaking Tests?	4.22
How well do you think the Interlocutor frames work?	4.33
In your experience, how positive is the washback on the classroom?	4.31
How well prepared do YLE candidates in your country/region tend to be?	4.24
How suitable is the content and format of the Speaking Test materials for the cultural backgrounds of the YLE candidates in your country/region?	3.83
How satisfied are you with the current training materials?	4.59

Practical Issues

A number of practical issues were also raised during the consultation process. For example, some centres have mentioned that the glossy paper of the test booklets causes problems, particularly in the colouring task of the Listening paper. We have now replaced the glossy paper with normal matt paper and this change has been positively received.

Cambridge ESOL is aware of the wide-ranging impact on teachers, children, parents and publishers that any changes that we make to a test can have. We therefore always give test users at least two years' notice of any changes to the tests and make information available through our website, printed materials and teachers seminars.

Future developments

Data are currently being analysed and the findings will influence the next phase of trialling. We will report on all these findings in a future issue of *Research Notes*.

We would like to take this opportunity to thank all the centres and individuals who assisted us with the first round of trialling for the YLE revision. We will soon be contacting centres to assist us with the second round. If you are able to help us, we will be very happy to hear from you. Please contact:

Trish Burrow, Subject Officer YLE (burrow.t@ucles.org.uk)

Juliet Wilson, Subject Manager YLE (wilson.j@ucles.org.uk)

References and further reading

Ball, F (2001): Investigating the YLE story-telling task, *Research Notes 10*, 16–18, Cambridge: UCLES.

Saville, N (2001): Test development and revision, *Research Notes 4*, 5–8, Cambridge: UCLES.

Creating a virtual community of assessment practice: towards 'on-line' examiner reliability

STUART D SHAW, RESEARCH AND VALIDATION GROUP

Introduction

Despite a considerable body of research on the reliability of examiners' marking in EFL (Alderson et al. 1995; Lumley and McNamara 1995; Weigle 1994, 1998), comparatively little research has been conducted on the procedures which may improve marking reliability. Only recently has a theory of marking reliability been proposed: reliable marking is purported to be produced by having an effective *Community of Practice*.

Community of Practice is a theory which describes how learning and professional practice is both a social and a cognitive enterprise. The term is a synonym for group, team or network – the phrase was coined by social science researchers who have studied the ways in which people naturally work and play together. In essence, communities of practice are groups of people who share similar goals and interests. In pursuit of these goals and interests, they employ common practices, work with the same tools and express themselves in a common language. Through such common activity, they come to hold similar beliefs and value systems.

Research literature about assessment and Communities of Practice suggests that a tight network or team and communication between examiners can facilitate the reliability of writing assessment. Wolf (1995:77) has suggested that:

> *marker reliability is lower the less the markers concerned form part of a group in constant contact and discussion with each other.*

Konrad (1998:5) has also argued that the reliability of assessment could be improved by the introduction of a Community of Practice. Hall and Harding (2002) have coined the phrase a '*Community of Assessment Practice*' in their investigation of whether communities of assessment practice exist in UK primary schools for the purposes of facilitating the consistent application of assessment criteria from the National Curriculum in English. They have suggested that a community of assessment practice improves examiner consistency and, perhaps more importantly, accuracy.

This article provides a brief review of Community of Practice in the context of second language writing assessment and considers the possible development of virtual examiner communities of Cambridge Writing examiners using electronic, asynchronous (delayed response time) communication.

Examiner communities and e-mail discussion groups

An increasingly important and significant aspect of technology in education is Computer-Mediated Communication (CMC) which includes video-conferencing, 'chat rooms', web-based bulletin boards and e-mail discussion lists. These technologies use computers to facilitate communication between people or groups of people. They are e-mail based and work by connecting together people who share a specific interest in a particular topic.

Cambridge ESOL has a strong interest in the potential of technology to improve communication between examiners within the examining process. Email discussion lists (E-lists), in particular, have the potential to foster conditions in which collegiate, reflective, practice-based development can occur, allowing examiners to share experience, information and good practice (Lieberman 2000). E-lists comprise a set of features that make them especially suitable for building examiner communities. They:

- support *many-to-many* communication, facilitating inter-group communication;
- are *asynchronous* – participants are not required to be 'on-line' simultaneously;
- are 'push' rather than 'pull' – the information comes to the user rather than the user having to retrieve it;
- are *text-based*- text allows structured discussion and, in conjunction with asynchronicity, can promote reflection which can lead to richer contributions;
- allow the creation of *searchable archives* such that messages and discussions can be retained for future reference and research.

UCLES have already demonstrated that E-lists for Media Studies and Psychology have successfully developed an electronic or virtual Community of Practice to support teachers and examiners (Riding 2002).

Communities of Practice in second language writing assessment

A review of the writing assessment literature shows that there is evidence to suggest that close contact and discussion between assessors and using exemplars of candidates' work facilitates the communication of standards and facilitates reliability. Wolf (1995) argues that in assessment systems the use of examples of candidates' work is particularly important as the standard is illustrated by the candidates' work rather than by descriptions of their work or by written assessment criteria or indicators. The reason for this is that if assessment criteria are separated from students' work they could be interpreted as appropriate for many different levels of achievement.

Evidence suggests that examiners value the opportunity in the co-ordination meeting to develop a 'community of assessment practice' and learn about the application of the mark scheme (Baird et al. 2002). The co-ordination of examiners involves examiners learning the standard of work which is required by a banded mark scheme. The co-ordination meeting is useful as it gives examiners a feeling of being part of a team, boosts confidence and provides examiners with feedback. At co-ordination meetings, senior or experienced examiners and new examiners tend to negotiate the interpretation of writing standards.

Communities of practice facilitate learning with the result that more experienced members of the group pass the social practice on to newcomers who in turn might contribute new ideas to the social practice of the group. In this way members of all levels of experience have some ownership of the social practice and the social practice itself is dynamic (Wenger 1998). The Community of Practice literature fits well with literature about standards as Cresswell (2000) explains when he states that standards are socially constructed and that applying standards is a form of social practice.

True ownership of a marking scheme, it would seem, leads to more accurate marking and a feeling of joint ownership amongst a team of markers leads to greater inter/intra-rater reliability (Barrett 2000).

Wenger (1998) explains that there are different forms of participation within the community and there are generations of members and the membership of the community is continually shifting. A community of practice can be sustained as long as there are generation relationships between newcomers and old-timers. A community of practice might not congeal when their assignment is started and it might continue when the assignment is finished. So the community of examiners (and the standard of examinations) can be maintained from one examination session to another as long as there are some old-timers in the group.

Lave and Wenger (1998) describe the process of newcomers being included in a community of practice. The newcomer must be given enough legitimacy to be accepted as a potential member of the community. Without this legitimacy the newcomer will have a hard time learning. Newcomers must transform their experience until they achieve the competence defined by a community. Old-timers too need to adapt as practice evolves. Wenger (1998) argues that one of the general guidelines for developing a community of practice is that learning is a process of participation, whether for newcomers or old-timers.

In the context of Cambridge ESOL examining, all examiners must conform to the mark scheme as defined by the Principal Examiner (or community) whether they are newcomers who might align their teaching experience with the community of practice or experienced examiners who might learn to assess new forms of knowledge and skills. A community of examiners might last for one examination session or a number of sessions. Over time Assistant

Examiners might become Team Leaders and perhaps Principal or Chief Examiners responsible for training new Assistant Examiners. Within a community of examiners, old-timers need to adapt as practice evolves.

Creating a virtual community of CELS writing examiners

In the light of the above, Cambridge ESOL are investigating the development of an on-line community of CELS writing examiners using e-mail discussion lists with a view to using electronic or virtual communities to promote consistency of assessment. The investigation will focus on the development of a Community of Practice amongst writing examiners and will address a number of specific research concerns:

- Does an e-mail discussion list for CELS Writing examiners create a Community of Practice?
- If it does, how is that Community formed?
- Can the reliability of CELS Writing assessment be improved by developing a Community of Practice?

The discussion list is for both Examiners and selected Cambridge ESOL staff, some of whom act as list administrators. The list – initiated in October 2003 – will be allowed to run for one year. The messages will provide evidence or otherwise of the development of a Community of Practice.

CELS examiners are invited to 'subscribe' to the email discussion list. Any e-mail that they then send to the list is automatically forwarded to everyone else who has subscribed. Since the lists are e-mail-based, subscribers do not need to be on-line at the same time, as the mail will wait in their in-box until they have the opportunity to read them. Messages are also archived on a special website that develops into a valuable library for members. Special software facilitates the management of these lists and sends out welcome messages to new members, helps control access to the communities and maintains the archive. It is hoped that the list will:

- generate lively debate and foster the conditions in which CELS examiners can share experience, information and good practice;
- promote professional development which should, among other things be ongoing and collaborative; include opportunities for *individual reflection* and *group enquiry*; be rooted in the *knowledge base* of examining; and be *accessible* and *inclusive*.

Analysis of the messages on the discussion list should show how a virtual community of practice is formed amongst practitioners and the benefits of such a community. Wenger (1998) provides a theoretical framework in which he lists 14 indicators that a Community of Practice has been formed. Messages on the list will be assessed against these indicators to establish whether such a Community of Practice has in fact been established. A questionnaire will be sent to the examiners in order to ascertain their feelings towards such an innovation including, for example, the list's advantages and disadvantages.

Conclusion

The information provided by the creation of an 'on-line assessment community' will be useful in the future to inform the developments of other virtual Communities of Practice, especially in relation to *Electronic Script Management* (ESM). Such communities might facilitate the reliability of marking if they were utilised in the standardisation process of electronic marking. This would be achieved by examiners posting queries on a secure discussion website. Team Leaders and Principal Examiners would be able to answer the queries to the benefit of all examiners as all examiners for the Writing paper would be able to read the website.

References and further reading

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

Baird, J, Bell, J F and Greatorex, J (2002): 'Tools for the trade': What makes GCSE marking reliable? *Learning Communities and Assessment Cultures: Connecting Research with Practice*, 28–30 August 2002, University of Northumbria UK.

Barret, S (2000): HECS LOTTO: *Does Marker Variability make examinations a lottery?* Division of Business and Enterprise, University of South Australia. www.aare.edu.au/99pap/bar99789.htm

Cresswell, M J (2000): The Role of Public Examinations in Defining and Monitoring Standards, in Goldstein, H and Heath, A (2000): *Educational Standards*, New York: Oxford University Press.

Hall, K and Harding, A (2002): Level descriptions and Teacher Assessment in England: towards a community of assessment practice, *Educational Research*, 44/1, 1–16.

Konrad, J (1998): *Assessment and Verification of National Vocational Qualifications: a European quality perspective*, Education-line www.leeds.ac.uk/educol/index.html.

Lave, J and Wenger, E (1991): *Situated Learning Legitimate Peripheral Participation*, Cambridge: Cambridge University Press.

Lieberman, A (2000): Networks as Learning Communities – Shaping the Future of Teacher Development. *Journal of Teacher Education*, 51/3, 221–227.

Lumley, T E and McNamara, T F (1995): Rater characteristics and rater bias: implications for training, *Language Testing* 12/1, 54–71.

Riding, P (2002): On-line Teacher Communities and Continuing Professional Development, *Teacher Development*, 5/3, 283–295.

Wenger, E (1998): *Communities of Practice Learning, meaning and identity*, Cambridge: Cambridge University Press.

Wolf, A (1995): *Competence Based Assessment*, Milton Keynes: Open University Press.

Reliability in First Certificate in English objective papers

ARDESHIR GERANPAYEH, RESEARCH AND VALIDATION GROUP

Introduction

Reliability is generally accepted to be an important aspect of test quality. On the one hand, reliability is a statistic – a number which looks more impressive the closer it gets to 1. The common statement that *reliability refers to the results obtained with an evaluation instrument, not to the instrument itself* is a statement about reliability as a statistic. At the same time, reliability clearly relates to the tendency of a set of test items to define a single internally consistent, uni-dimensional trait. In this sense reliability is actually akin to construct validity (at least, if the trait it defines is the same as the one the test claims to measure).

This article discusses reliability in language tests and presents some performance data for FCE objective papers: Reading, Listening and Use of English.

Reliability in language tests

There are two particular features of language tests that are important to a discussion of reliability:

Language tests fit into a framework

There has been a great deal of interest in recent years in the development of frameworks for language proficiency (the ALTE Framework and the Council of Europe Common Framework are but two examples). Depending on purpose, language tests may cover an entire system of levels (e.g. for placement) or focus on one particular level (as in achievement testing, or in suites of proficiency exams like the Cambridge Main Suite). What is important is that performance in both types of test should be related to the overarching framework, with as much precision as the purpose of the test allows.

Where exams test language proficiency at a single level, within a comprehensive system of levels, they can be expected to provide more efficient, focused, relevant, and in-depth measurement of ability. However, candidates of widely-ranging ability are easier to rank reliably, and so it is not uncommon for short placement tests (i.e. QPT) to show reliability indices as high or higher than extended, multi-component exams at a single level (i.e. CPE). Clearly, the problem here is with the notion of reliability, rather than the measurement properties of the exam itself. Other things being equal, a test which focuses on a narrow range will provide more information about candidates in that range than will a similar-length test with a wide focus. This is precisely the rationale for computer-adaptive testing, which automatically adjusts task difficulty to match the estimated ability of individual candidates.

Communicative language tests cover a range of skills

At both the level of the individual component and of the whole exam, reliability, in the sense of internal consistency, may present problems for attempts to develop language tests modelled on a theory of communicative language ability.

> *The plausibility of internal consistency estimates appears to be further compromised by the deliberate efforts made to introduce variety and heterogeneity into examinations. If the principle of inclusion in an examination is to mix modalities and skill requirements and contents so that individual opportunities to respond well are enhanced, why expect internal consistency? (Wood 1993:138)*

As this suggests, communicative tests may be heterogeneous in two ways:

- The tasks tap a broad range of language skills;
- The candidates bring very different profiles of skill to bear, which may be taken to represent equally valid expressions of ability.

At component level, task-based exercises have been replacing discrete point multiple choice items in communicatively-oriented exams in order to provide greater context and authenticity (both situational and interactional). In consequence, they may tap a broader range of language skills than a narrowly-conceived, psychometric test. An additional consequence may be that items take longer to respond to, and so fewer can be accommodated within practical time constraints. This may bring about a small reduction in the estimated reliability using an internal consistency estimate when compared with tests using a greater number of discrete point items.

At exam level, the components set out to measure distinct aspects of language proficiency, hence the intercorrelations are generally not high. The composite reliability of the exam is still considerably higher than any individual component, because of the large number of observations (marks) which are combined. However, the heterogeneity sets practical limits to the possible composite reliability of the exam.

Reporting reliability

There are two major issues with reliability and language exams:

1. Reliability, as a statistic, is defined as an interaction between a specific test and group of respondents, and thus has an inherently local, limited meaning. Language tests need interpretations, which relate to larger frames of reference.

2. Reliability, as the internal consistency of a test, may conflict with the heterogeneous, inclusive conception of communicative language ability.

The overall reliability of an exam such as FCE comprising several component papers is known as its composite reliability. Just as the reliability of individual components, at least when using internal consistency estimates, depends on the homogeneity of the test items, so the composite reliability reflects the intercorrelation of the components. The stronger the intercorrelation, the higher the composite reliability. Cambridge ESOL uses the Feldt and Brennan (1989) method for estimating composite reliability.

The composite reliability of the final grade covers all the papers within one exam and is usually higher than the reliability of individual papers. The composite reliability of the FCE exam has consistently been in the range of 0.92 since June 2000.

In addition to the composite reliability, the reliability of each paper needs to be reported. In this issue we report the reliability estimates for the three objective papers in FCE: Paper 1 (Reading), Paper 3 (Use of English) and Paper 4 (Listening). There are as many as three different exam versions for each Reading and Use of English Paper during December and June sessions. There are more versions for the Listening Paper at each session because of the physical exam constraints and the security of the paper. For ease of reference we report only the reliability of the two common listening papers that are used across the world.

Table 1: Reliability estimates for FCE Paper 1 2000–3

Version	Session	Year	No. Items	Alpha	Sample Size[1]
1	December	2000	35	0.82	34828
2	December	2000	35	0.79	24400
1	March	2001	35	0.85	14343
1	June	2001	35	0.86	35876
2	June	2001	35	0.79	43963
1	December	2001	35	0.82	33756
2	December	2001	35	0.85	28271
1	March	2002	35	0.84	17110
1	June	2002	35	0.85	35012
2	June	2002	35	0.83	59107
3	June	2002	35	0.84	48851
1	December	2002	35	0.85	30922
2	December	2002	35	0.83	34582
1	March	2003	35	0.84	18963
1	June	2003	35	0.83	37107
2	June	2003	35	0.82	50907
3	June	2003	35	0.85	57325

Table 1 demonstrates that the Reading paper has shown an average reliability of 0.84 in the year 2002/2003. This figure is quite respectable for a test of 35 items, all of which are based on communicative tasks. A similar picture can be seen in Table 2 for the Listening paper where there are only 30 items in the test.

Table 2: Reliability estimates for FCE Paper 4 2002–3

Version	Session	Year	No. Items	Alpha	Sample Size[1]
1	June	2002	30	0.81	54959
2	June	2002	30	0.83	54560
1	December	2002	30	0.86	36858
2	December	2002	30	0.84	35070
1	June	2003	30	0.87	54658
2	June	2003	30	0.85	54497

Table 3 shows the reliability estimates for the Use of English paper.

Table 3: Reliability estimates for FCE Paper 3 2000–3

Syllabus	Session	Year	No. Items	Alpha	Sample Size[1]
1	December	2000	65	0.89	35438
2	December	2000	65	0.88	25476
3	December	2000	65	0.88	36964
1	March	2001	65	0.91	15133
1	June	2001	65	0.91	38128
2	June	2001	65	0.92	60640
3	June	2001	65	0.90	48198
1	December	2001	65	0.91	34598
2	December	2001	65	0.92	36690
3	December	2001	65	0.89	28814
1	March	2002	65	0.92	17413
1	June	2002	65	0.91	35511
2	June	2002	65	0.93	58104
3	June	2002	65	0.91	49785
1	December	2002	65	0.90	29829
2	December	2002	65	0.90	34252
3	December	2002	65	0.91	30843
1	March	2003	65	0.92	19312
1	June	2003	65	0.92	36666
2	June	2003	65	0.91	50499
3	June	2003	65	0.91	56237

Table 3 shows that on average the reliability of Paper 3 has been in the region of 0.91 since 2001. The relatively higher reliability estimate of this paper may be due to the number of items in the test, which has a direct relationship with the increase in Alpha.

Conclusion

To recapitulate, we have explained that a single level exam is likely to have a lower reliability estimate than a multilevel exam. However, it seems from the above evidence that the average reliability figures reported for FCE objective papers are acceptable for these tests with the given number of items and the restricted range of the population ability taking these papers. We have also argued that communicative task-oriented tests add an additional

[1] The actual candidate entries are higher than the samples reported here.

restriction on reliability estimates. The more complex the tasks are, the less likely it is that we achieve high Alphas; this is due to the wide focus of the task levels used. The figures reported in this article, nevertheless, illustrate that despite these factors FCE reliability estimates tend to be very high and consistent across sessions.

Reference

Feldt, L S and Brennan, R L (1989): Reliability, in Linn (ed): *Educational Measurement: Third Edition*, American Council on Education, New York: Macmillan.

Wood, R (1993): *Assessment and Testing*, Cambridge: Cambridge University Press.

Announcement of the winner of the IELTS Master's Award 2003

As part of the tenth anniversary of IELTS in 1999, the IELTS partners – University of Cambridge ESOL Examinations, The British Council, and IDP Education Australia – agreed to sponsor an annual award of £1000 for the master's level dissertation or thesis which makes the most significant contribution to the field of language testing.

For the award in 2003, submissions were accepted for dissertations completed in 2002. The IELTS Research Committee, which comprises members from the three partner organisations, met in November 2003 to review the shortlisted submissions. The Committee was extremely impressed with the high quality of all the shortlisted dissertations and it was agreed that the candidates and their work are indicative of the considerable worldwide interest in language testing and merit a wider audience.

After careful consideration, the Committee decided to announce one winner: Eunice Eunhee Jang – *In search of folk fairness in language testing.*

Singular in its methodology and intent, the Committee considered the winning dissertation to be a very impressive piece of work with several key points of resonant originality. The dissertation investigated ESL test takers' perception of fairness from two strikingly different methodological perspectives: a sophisticated survey instrument as well as follow-up qualitative data. Coining the phrase 'folk fairness', which is new to both language testing and general studies of educational and psychological measurement, the term is itself meritorious of a separately publishable paper. The topic, original perspective and extent of the study mean this is an important work deserving a wider readership.

The dissertation was a large scale study which had been very carefully planned and which was well presented. The Committee believed that the writer had a very clear understanding and awareness of the issues and that the work was well beyond the level expected from most MA students.

The abstract from the award-winning dissertation is presented below:

> *Fairness is generally defined as a social value that applies in understanding, interpreting social actions, and making decisions over human performance. In the field of testing, the concept of fairness is commonly defined as social justice or equity which is associated with equal opportunity and equitable treatment. The purpose of the present study is to seek to understand the process through which examinees construct and interpret the concept of fairness in the standardised language testing such as the Test of English as a Foreign Language (TOEFL).*
>
> *To investigate three main research questions proposed in this study, I utilised a mixed methods research design comprising both quantitative and qualitative approaches.*
>
> *Reliabilities and exploratory factor analyses were performed to examine the factor structure of the concept of test fairness using large-scale quantitative data. Several Analysis of Variance tests were performed to examine group characteristics effects on conceptions of test fairness. Focus group and individual interviews were conducted to understand test takers' making sense of their test taking experience in relation to test fairness.*
>
> *Results from analyses of the data from both quantitative and qualitative research methods support the conclusion that the concept of test fairness is multi-faceted, dynamic and both culturally and contextually situated.*
>
> *The author argues that fairness is not derived from a test itself, but it is constructed, interpreted and practiced in different ways by various stakeholders. Therefore, the goal of test fairness studies should not be only to seek for a decisive criterion by which a test is judged as either 'fair' or 'unfair', but also to represent diverse voices of co-participants in testing practice. The study concludes that social inquiry about test fairness would be enhanced through collaborative dialogic engagement with all testing practice participants.*

The presentation of the 2003 IELTS Master's Award will be reported in a future issue of *Research Notes*. If you are interested in submitting your Masters thesis or dissertation for the 2004 award, please read the following information.

IELTS Master’s Award 2004

For 2004, the entry procedures and timetable for the award are given below:

Submission and evaluation procedures

Dissertations will only be considered eligible if they were submitted and approved by your university in 2003. Dissertations completed in 2004 will not be considered eligible for the 2004 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 dissertation should be language testing focused but need not be IELTS-related.

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
ESOL Division
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 submissions 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 2004:

1 June	Deadline for submission of dissertation extracts and supervisor’s reference to University of Cambridge ESOL Examinations
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

Details of the application process for the IELTS Master’s Award 2004 can also be found on the IELTS website – www.ielts.org Please note that submission details may change from year to year and it is therefore important that the most current procedures are consulted.