

ResearchNotes

Contents

Editorial Notes	1
Second language writing assessment: Cambridge ESOL's ongoing research agenda	2
IELTS writing: revising assessment criteria and scales (Phase 3)	3
Exploring the relationship between YLE Starters and Movers and Breakthrough level	7
BAAL/CUP Workshop-Vocabulary Knowledge and Use: Measurements and Applications	8
Making the grade: score gains on the IELTS Writing test	9
Question uptake in the Certificate in Advanced English Writing Paper	14
ESOL Staff Seminar programme	15
Conference Reports	18
Other News	20
IELTS joint-funded research 2004/5 (Round 10): call for proposals	22
Award presentations at LTRC 2004	24

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

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

Editorial Notes

Welcome to issue 16 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

This issue focuses on the productive skill of Writing which is an integral component for most of our language testing products, with the exception of modular products such as CELS and BULATS in which candidates can opt to take the Writing component as they wish. This issue considers a range of issues linked to assessing learners' proficiency in writing including general assessment issues and more specific issues such as how we develop and use rating scales to accurately assess writing and how exam levels can be equated by investigating vocabulary and what is included in each syllabus.

In the opening article Lynda Taylor considers some of the research issues relating to second language writing assessment within Cambridge ESOL. In the following three articles we consider different aspects of writing test development and validation. Firstly Stuart Shaw describes our activities in revising the IELTS rating scales for assessing writing. Although concentrating on IELTS, Stuart's article has relevance for other ongoing rating scale revisions for multiple components, not just writing. In the following article Trish Burrow reports on a recent study that explored the relationship between the content coverage of Starters and Movers syllabuses and the Council of Europe's Breakthrough level, looking particularly at vocabulary and test syllabuses. Whilst not primarily focussing on writing, this article is relevant to all of our examinations.

Returning to IELTS, Tony Green outlines a range of research projects into the extent and predictability of IELTS writing score gains. This has relevance for many IELTS candidates for gaining admission to universities or other purposes. The correct measurement of Writing proficiency is of equal importance to all exams, whether high stakes like IELTS or low stakes like YLE. Finishing our focus on writing we present some data on the uptake of optional questions in the CAE Writing paper in the second of our series on performance data.

We then review some of the other activities that Research and Validation staff take part in, in the form of the professional development through staff seminars and conference attendance, many of which focus on the productive skills of writing and speaking. Rowena Akinyemi and Fiona Barker review some of the contributions to the 2003 staff seminar programme made by external speakers. In our Other News section we list the articles from issues 7–15 which can be downloaded direct from the *Research Notes* website.

We end this issue with details of how to apply for the tenth IELTS joint-funded research programme for 2004/5 and photographs of two award presentations which took place at the 2004 Language Testing Research Colloquium in Temecula, USA.

Second language writing assessment: Cambridge ESOL's ongoing research agenda

LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

A direct test of writing has been an integral component of most of our Cambridge ESOL examinations ever since the Certificate of Proficiency in English (CPE) was introduced in 1913. At that time the inclusion of a written 'essay' or 'composition' in Cambridge's EFL examinations probably reflected the longstanding tradition in British education of academic assessment via written examination; later in the 20th century, however, the inclusion of a direct test of writing reflected the growing interest in communicative language ability and the importance of performance-focused assessment.

Over the past ten years, there has been increasing recognition that direct assessment of L2 writing ability involves a complex interaction between different variables or facets (Milanovic & Saville 1996); since the early 1990s Cambridge ESOL has focused considerable research effort towards improving our understanding of some of the key variables involved in writing assessment and the way in which they interact with one another.

The nature of the *writing ability construct* has been a major focus of attention. Our work to develop a Common Scale for Writing dates back to 1994 when Phase 1 set out to analyse the relationship between levels of the Cambridge Main Suite writing tests (KET to CPE) and to hypothesise a common scale of writing ability. Phase 2 of the project (1999–2004) has been examining distinguishing features of writing performance at different ability levels, especially in terms of their domain specificity (i.e. general, academic, workplace). Much of this work has been reported in previous issues of *Research Notes* and a fuller article will shortly appear in the journal *Assessing Writing* (Hawkey and Barker forthcoming).

Investigation into the *behaviour of writing examiners* has been another major strand of our research agenda for writing assessment. Early studies explored the marking strategies and decision-making behaviour of Cambridge EFL writing examiners for FCE (Milanovic, Saville & Shuhong 1996), but over time investigation has extended to other ESOL tests, including CPE and IELTS. Improved understanding of rater behaviour has had a positive impact on our procedures for *examiner training and standardisation*; such procedures are generally accepted as being critical to reliability and validity in second language performance assessment.

Another key focus of our attention has been the study of writing *assessment criteria* and the *development of rating scales* for assessing writing performance. This has proved especially important in the context of revising our existing Writing tests

(e.g. CPE, KET/PET, IELTS) and developing new ones (e.g. CELS). Analysis of actual samples of writing performance has always been instrumental in helping us to understand more about key features of writing ability across different proficiency levels and within different domains. The relatively new field of *corpus linguistics* offers considerable promise in this area and our exploitation of the Cambridge Learner Corpus is beginning to deliver fresh and exciting insights into the nature of learner written performance in terms of its grammatical, lexical and discursive features.

Technological advances continue to offer assessment providers new and innovative opportunities for testing writing proficiency and for some time now Cambridge ESOL has been undertaking studies in the following key areas: the relationship between computer-based and paper-and-pencil tests of writing (e.g. BULATS, IELTS); the use of automated grading systems for writing assessment; the use of electronic script management systems; the role of electronic discussion lists in developing a community of assessment practice (CELS).

The last few years have seen a much clearer identification and understanding of the many different avenues which are open to researchers in the field of writing assessment. The development of new and sophisticated qualitative and quantitative methods to facilitate our investigative and analytical research work has been equally important over this period; advances in discourse analysis, verbal protocol analysis and in multi-faceted Rasch measurement are just a few examples. Work in all of the areas highlighted above is regularly profiled in *Research Notes* and at conferences and seminars. In future we hope to be able to disseminate more of it in the public domain through articles in refereed journals and through a series of published volumes.

References and further reading

- Hawkey, R and Barker, F (forthcoming) Developing a common scale for the assessment of writing, *Assessing Writing*.
- Milanovic, M and Saville, N (1994) *An investigation of marking strategies using verbal protocols*, Cambridge: UCLES EFL Internal Report.
- Milanovic, M, Saville, N and Shuhong, S (1996) A study of the decision-making behaviour of composition markers, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Cambridge: Cambridge University Press, 92–114.

IELTS Writing: revising assessment criteria and scales (Phase 3)

STUART D SHAW, RESEARCH AND VALIDATION GROUP

Background

Routine monitoring of operational test performance together with ongoing validation studies sometimes leads to a decision to revise certain features of the test in question, e.g. test format, task design, assessment criteria or rating scales; revision of any of these features requires that appropriate validation studies are conducted before the revised test is implemented in the live operational context.

The initial phase of the revision of assessment criteria and rating scale descriptors for the IELTS Writing Modules – the *Consultation, Initial Planning and Design* Phase – was completed in December 2001 (reported in *Research Notes* 9). Phase 2 of the project (*Research Notes* 10) – the *Development* Phase – entailed the design and development of the revised rating scale; this work was completed in May 2003 in preparation for trialling and validation.

It is essential to the success of the revision project that considerable effort is devoted to the validation of the rating scale prior to its widespread use. To this end, a detailed and thorough validation programme – employing both qualitative and quantitative methods for the establishment of validity, reliability, impact and practicality (VRIP) – has been undertaken. In revising the IELTS rating scale an attempt was made to achieve an optimum balance among the VRIP qualities. Successful validation of the IELTS revised assessment criteria and band level descriptors is dependent upon all VRIP features being dealt with adequately and completely.

The quantitative and qualitative dimension of the third phase of the project – the *Validation* Phase – began in July 2003; it has involved the collection and subsequent analysis of data, in the form of examiner scores and questionnaire responses, from a multiple rating study conducted simultaneously in the UK and Australia.

Developing and using scales – some assumptions

In general, rating scales attempt to equate candidate performance to specific verbal descriptions (Upshur & Turner 1995). The development (and subsequent revision) of a rating scale and descriptors for each scale level are of great importance for the validity of any assessment (Weigle 2002). Raters are expected to make decisions on the basis of common interpretations of the scale contents. This process should be transparent and simple (Pollitt & Murray 1996, Zhang 1998).

It is widely recognised in the assessment reliability literature that the shared interpretation of rating scale descriptors cannot be assumed; unless rating scale points define clearly differentiated levels or bands (Bachman 1990), precise interpretation by different audiences is likely to vary and will do so according to ‘previous experience, unconscious expectations and subjective preferences regarding the relative importance of different communicative criteria’ (Brindley 1998:63).

Successful performance assessment depends upon sound examiner judgements; if we can create a rating scale (or set of scales) that describe instances of written performance in a valid way, *and* if we can train examiners to understand and assimilate its contents, then the scale will be used validly and reliably.

Issues raised by the IELTS rating scale revision

In considering the nature of the rating process it is necessary to understand how the revised rating scale is actually going to be used by IELTS examiners, as well as the role of any future re-training of those examiners. Some of the issues related to the revised scale include:

- Does the scale capture the essential qualities of the written performance?
- Do the abilities the scale describes progress in the ways it suggests?
- Can raters agree on their understanding of the descriptions that define the levels?
- Can raters distinguish all the band levels clearly and interpret them consistently?
- Can raters interpret effectively any ‘relative’ language terms e.g. ‘limited’, ‘reasonable’, ‘adequate’?
- Do raters always confine themselves exclusively to the context of the scale?
- What is the role of re-training IELTS examiners in the use of the new rating scale in the rating process?

A carefully designed multiple rating study can provide answers to these and similar questions and can help confirm that the rating scale and descriptors are functioning as intended before they are used operationally.

Research Design

In total, fifteen raters participated in the multiple rating study – 3 Trainers (2 from the UK and 1 from Australia) and 12 Senior

Examiners (4 from the UK and 8 from Australia). The trainers were fully familiar with the revised rating scale having been instrumental in its development. The remaining participating raters were all highly experienced IELTS examiners and were identified as representative of the worldwide population – or ‘universe’ – of examiners for the test.

The scripts used in the study were all sample monitoring scripts. 60 scripts (30 General Training Writing and 30 Academic Writing) were selected with 30 of each task (Task 1 and Task 2) across all band levels. The 60 writing performances were identified as benchmark scripts representing the full proficiency continuum for the test.

In outline, the procedure was to standardise the group of trial examiners using the revised rating scale, do multiple marking of a set of scripts, and then, off-site, do a series of iterations in which further sets of scripts are marked following further standardisation exercises.

Data collection took place on four occasions over one month. The apportionment of scripts and the timing of project iterations in relation to stages of the training process are tabulated in Table 1.

Table 1: Timetable for data collection and script apportionment

Standardising Exercise	Timing	Number of Scripts	Batch Number
Initial marking following introduction to revised rating scale (IT1)	September 25th, 2003	8 (+ 2 ‘standard’ setting scripts)	1
Standardisation Training (IT2)	September 25th, 2003	10	2
Iteration (IT3)	2 weeks after initial training	20	3
Iteration (IT4)	4 weeks after initial training	20	4
Total	4 weeks	58	

The first iteration (IT1) immediately preceded standardisation training; the second (IT2) immediately followed standardisation training. IT3 and IT4 were carried out approximately two and four weeks respectively after training. The full procedure is shown in diagrammatic form in Figure 1 and consisted of three phases:

- a) Preparation of scripts for use in standardisation
 - Scripts were triple marked (3 x trainer ratings) to provide benchmark ratings;
 - A commentary was provided for each script explaining why the given mark was correct;
 - All annotations were erased from the scripts;
 - Sufficient copies were made for each trial examiner i.e. 60 scripts x 12 examiners = 720 copies.
- b) Examiner training day: full-day, on-site induction to the study and initial standardisation training (September 25th 2003)

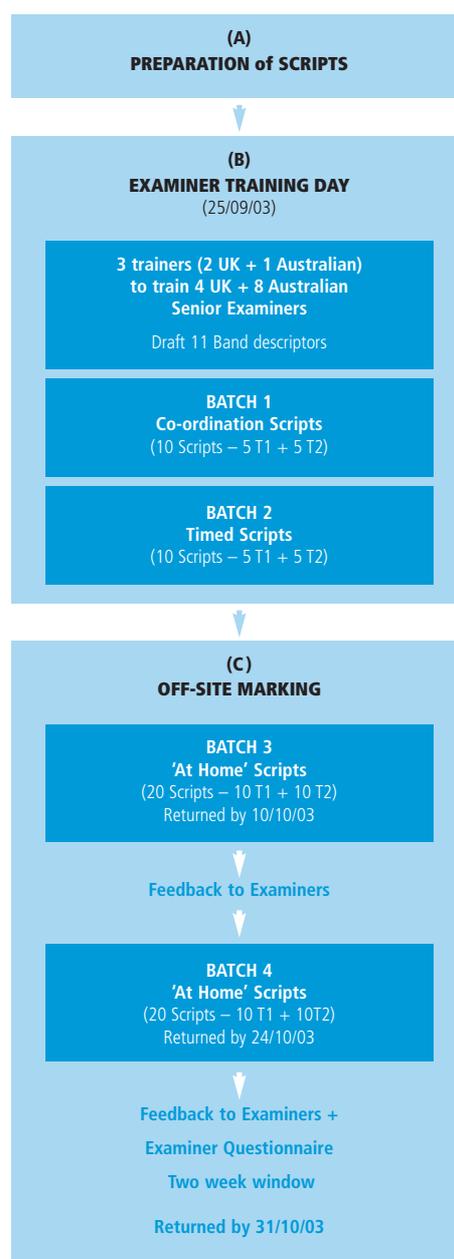


Figure 1: Procedure for data collection

- The three trainers and all 12 examiners (4 UK + 8 Australian) were presented with the primary objectives of the study and were given an outline of the methodology;
 - Initial marking session was conducted on Batch 1 scripts, i.e. a total of 8, using the revised draft descriptors;
 - A second marking session was conducted following standardisation training using Batch 2 scripts, i.e. a total of 10;
 - Batch 3 scripts, comprising 20 scripts were handed out at the end of the meeting to be marked at home.
- c) Off-site marking and continued standardisation over a period of approximately four weeks:
 - Further standardisation in an iterative manner was conducted (involving 2 iterations – Batches 3 + 4);

- At each iteration, the trial examiners marked further standardisation scripts (20 per batch) and checked their performance after each script using the explanatory notes. Trial examiners only received feedback and marks for a batch of scripts after they had returned the marks for that batch.
- Retrospective data was captured by an examiner questionnaire.

Data Analysis and Results

The main descriptive statistics used to describe the distribution of marks given by the trial raters were the *mean*, or average mark (measure of harshness or leniency of the rating scale), the *standard deviation* or the average amount that marks differ from the mean (measure of use of available rating scale), and the *range* (measure of marks at each end of a distribution), which is the easiest way to talk about the spread of marks from the central. Mean scores, standard deviations and range were calculated by rater group, by batch and by subscale. Results suggest that, relative to the benchmark ratings, both UK and Australian raters tended to undermark slightly and over a narrower range. Raters tended to be most generous on the *Lexical Resource* criterion and most severe on the *Task Achievement/Task Response* criterion.

The scores given by the trial examiners were compared with the standard benchmark ratings for the same scripts, by subtracting the latter from the former. Thus, if an examiner gave a particular script a global score of 5, and the standard band score for that script was 5, the difference would be noted as zero; if an examiner gave a global score of 3 and the standard was 4, the difference was noted as -1 ; if an examiner gave a global score of 5 and the standard was 3, the difference was noted as $+2$, and so on. The frequency with which the difference was zero, or -1 , or $+3$, etc., was counted for each rater and for each iteration. Over 90% of ratings fell within \pm one band of the original benchmark rating. Mean trainer/rater intercorrelations for global scoring were consistently high – of the order of 0.9 and above; and trainer/rater intercorrelations for scoring on the analytical subscales showed similarly high values, ranging from 0.87 for Coherence and Cohesion to 0.92 for Grammatical Range and Accuracy.

Inter-rater reliabilities for multiple raters for each of the four batches were calculated by generating a Pearson correlation matrix. Any distortion inherent in using the Pearson for ordinal data was corrected for by applying a Fisher Z transformation to each correlation. This was done for the whole group of 12 and also for the 2 regional groups – UK (4) and AUS (8). Inter-rater reliabilities for the three examiner cohorts are shown in Table 2 and these remain encouragingly high for each rating group throughout the trial.

Strength of agreement between raters was calculated using the multi-rater version of Kappa which averages the proportion of agreement between raters before adjusting for chance agreement. Researchers vary on where to set the boundaries for Kappa values;

Table 2: Inter-rater reliabilities

	UK + Australia	UK	Australia
Batch 1	0.770	0.790	0.780
Batch 2	0.775	0.805	0.790
Batch 3	0.775	0.805	0.780
Batch 4	0.775	0.850	0.780

in this study values between 0.21–0.60 were considered as ‘moderate agreement’ and coefficients greater than 0.6 were regarded as ‘good’. The value achieved by the rater group at the end of Batch 4 was 0.45 and so fell into the ‘moderate’ range; they exhibited moderate to good levels of agreement in terms of their absolute ratings. Levels of agreement, however, did improve throughout the trial.

Generalisability and decision studies were carried out to estimate reliability of the rating procedure when applied by a single rater; this is the operational condition for the live writing test. Generalisability coefficients achieved throughout the study are shown in Table 3.

Table 3: Generalisability Matrix

	G Coefficient
Batch 1	0.846
Batch 2	0.838
Batch 3	0.932
Batch 4	0.878

The data from Batches 2 and 4 were subjected to a multi-faceted Rasch analysis and some of the output from this is shown in the vertical summary for Batch 4 in Figure 2. Working from left to right, we can see that the candidate abilities (scripts B4S1–20) cover a broad range -10.52 logits. The raters range only marginally in severity with the most severe raters at the top of the group (AUSmm and auspd). Calibration and fit statistics for the raters fell within acceptable limits. Even though the rater severity range was already narrow at the Batch 2 stage, it reduced substantially between Batch 2 and Batch 4, indicating that raters were becoming increasingly unified in their approach as they gained in experience. The most difficult scale is the *Task Achievement/Response* scale, the easiest is the *Lexical Resource* – with almost a logit between them; this is consistent with the earlier correlational evidence.

Summary of quantitative findings

The aim of this study was to evaluate the effectiveness of the revised criteria and scales for IELTS Writing. One indication of this

Figure 2: All Facet Vertical Summary

Measr	+script	-rater	-Scores	S.1
7				(9)
	B4S8_			---
6				8
	B4S12			---
5				7
	B4S6_			---
	B4S19			---
4				6
	B4S1_			---
	B4S16			---
3				5
	B4S17			---
	B4S13			---
	B4S3_			---
2				4
	B4S7_			---
	B4S20			---
1				3
	B4S14			---
	B4S4_			---
	B4S5_			---
0				2
	B4S18	AUSmm auspd	TA/R	---
	B4S15	UK4 ausrj	CC GRA	---
	B4S2	UK1 ausat auscp ausdo ausdp	LR	---
		GS_ UK2_ UK3_ auscm		---
-1				1
	B4S10			---
-2				0
	B4S9_			---
-3				(0)

-4				---
	B4S11			---
-5				---

is how well the raters agreed with the original benchmark ratings; over 90% of ratings were within +/- one band of the benchmark award. There was a tendency for raters to mark slightly more harshly and over a narrower range of the available markscheme; however, this finding is consistent with results from other studies (Weigle 2002) which suggest that 'new' raters (even those who are experienced but are working with a new or revised markscheme/rating scale) tend to mark harshly.

The correlational analyses suggested that trainers and raters shared a substantial amount of knowledge when rating at the assessment criterion level – evidence that both sets of examiners are 'singing from the same song sheet'. Greatest agreement existed for the *Lexical Resource* and *Grammatical Range and Accuracy* scales. Further correlational evidence supported the notion that raters were clearly distinguishing between the four assessment criteria. High inter-rater reliabilities were achieved – both at the beginning of the study and throughout. The strength of agreement analyses showed that although raters were in good general agreement on the rank ordering of scripts, they were in less agreement regarding the absolute mark assigned to those scripts – again this may point to level of exposure and experience.

Results from the G-theory studies suggest that raters use and interpret the markscheme in generally the same way. This was confirmed by the very good generalisability coefficients estimated for the single-rater condition which is the operational condition. The FACETS analysis was useful for identifying individual cases of misfit which, if investigated, might throw light on problems in interpreting the markscheme in its present form. The five analytical subscales do appear to measure distinct aspects of writing proficiency. *Task Achievement/Task Response* and *Lexical Resource*

were clearly distinct traits. The other two scales – *Coherence and Cohesion* and *Grammatical Range and Accuracy* appear to be more closely related. There is a small difference in the difficulty of the scales but they appear to contribute consistently to the candidate's final writing score. The FACETS graphical plot indicates that the band thresholds are fairly similarly spread in each of the scales. The results support the view of Lynch & McNamara (1998) that G-theory and FACETS are complementary approaches which both provide useful information.

In addition to confirming that criteria and scales are functioning well, results from the statistical analyses can inform future rater training. They will allow training to be targeted at achieving better consensus in the rating areas where the greatest discrepancies lie. Results are also helping to identify which writing scripts will be most reliable and useful for rater training/certification.

Qualitative insights

The quantitative nature of the multiple rating study was complemented with a qualitative data collection exercise using questionnaires and focus group discussion. Raters reported that they found the study to be a very positive experience; they considered the revised rating scale to be user-friendly and they welcomed the greater clarity in the new descriptors, believing them to provide a more comprehensive description of the key features of writing at each band level. The separation of *Lexical Resource* and *Grammatical Range and Accuracy* was perceived to be extremely valuable. *Coherence and Cohesion* was found to be the most difficult criterion to assess. Raters overwhelmingly agreed that the revised *Task Achievement* and *Task Response* subscales were effective for rating *General Training* and *Academic Task 1* and *Task 2* respectively. When rating they reported giving most attention to the *Task Achievement/Task Response* criteria, basing their assessments on a detailed study of the features of the candidate's response in relation to the task. Interestingly, in terms of productivity, raters were concerned that their overall rate of marking might be affected – at least initially: four (rather than the current three) criteria might elicit longer reading and processing time and ultimately lengthen the rating process. Raters understandably highlighted the inevitable implications for examiner fees.

Conclusions

Results of the quantitative analyses provide evidence of the validity and reliability of the revised assessment criteria and rating scales; supplementary findings from complementary qualitative studies point to the practicality and positive impact of the criteria and scales from the raters' perspective. Information from both quantitative and qualitative analyses is now feeding directly into the rater retraining and standardisation programme (the Implementation phase); details of this phase will be reported in a future *Research Notes*. The revised writing assessment criteria and scales for the IELTS Writing Test are on target to become operational from January 2005.

References and further reading

- Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.
- Brindley, G (1998) Describing Language Development? Rating Scales and SLA, in Bachman, L F and Cohen, A D (Eds) *Interfaces between Second Language Acquisition and Language Testing Research*, Cambridge: Cambridge University Press.
- Lumley, T (2000) *The process of the assessment of writing performance: the rater's perspective*, Unpublished PhD dissertation, Department of Linguistics and Applied Linguistics, The University of Melbourne.
- Lynch, B K and McNamara, T F (1998) Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants, *Language Testing* 15, 158–80.
- McNamara (1996) *Measuring Second Language Performance*. Harlow: Longman.
- Pollitt, A and Murray, N L (1996) What raters really pay attention to, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Cambridge: Cambridge University Press, 74–91.
- Upshur, J A and Turner, C E (1995) Constructing rating scales for second language tests, *ELT Journal* 49 (1).
- Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.
- Wright, B D and Stone, M H (1979) *Best Test Design*, Chicago, IL: MESA Press.

Exploring the relationship between YLE Starters and Movers and Breakthrough level

TRISH BURROW, EXAMINATIONS AND ASSESSMENT GROUP

As part of ongoing research at Cambridge ESOL, in early 2003 a consultant was asked to explore the relationship between the first two levels of the YLE Tests (Starters and Movers), and the Council of Europe Breakthrough level (Trim, in prep). The research mapped Starters and Movers to Breakthrough level in terms of content (topics and vocabulary) and also sought to answer the question “Are there features of the test syllabuses of Starters and/or Movers which could be usefully added to or influence Breakthrough?”

Mapping of content

In the first part of the research the lexical content of the Starters and Movers tests was mapped against topic headings for Breakthrough level. Table 1 shows which lexical items are shared between all three levels for a specific topic (*Family*) and which

- appear in Breakthrough level, but not Starters and/or Movers or
- appear either at Starters or Movers but not in Breakthrough level.

The research indicated that there are substantial areas of overlap between Breakthrough level and both Starters and Movers in terms of topic areas and lexis. It is clear from Table 1 that words appear exclusively in the ‘adult’ or ‘young learner’ columns; examples of the latter group of words include ‘grandmother’ and ‘grandfather’, whereas the Breakthrough syllabus includes words which have a clear relevance to adults e.g. ‘husband’ and ‘wife’. Breakthrough divides topics into four domains: *personal*, *public*, *vocational* and *educational*. Many of the topics which are listed within these domains are included in the syllabuses of Starters and Movers, although the focus may be on ‘school and the classroom’ (Starters) as opposed to the more abstract topic of ‘education’.

We can see from the following table how the topics in the

Table 1: Overlap of lexis for the topic ‘Family’

Breakthrough	Starters	Movers
Family family, father, mother, brother, sister, child (i) husband, wife, (ii) son, daughter	Family family, father, mother, brother, sister, child – – (iii) PLUS baby, grandmother, grandfather, dad, mum	(all in starters PLUS) son, daughter (iii) PLUS aunt, cousin, grandson, granddaughter, grandparent, parent, person/people, uncle

personal domain of Breakthrough level overlap in several places with topics in the Starters and/or Movers syllabuses.

Words which only appear in the Starters/Movers columns clearly have a relevance to children and not adults i.e. animals, classroom instructions and action verbs e.g. ‘kick’, ‘bounce’ etc. Words which appear in the Breakthrough column only and not in young learner syllabuses do so for several reasons. For example, children have either little concept of or motivation to use words such as ‘rent’, ‘insurance’, ‘memo’, and ‘to book’. This is particularly the case with younger children who have yet to grasp more abstract concepts.

The research report makes it easy for us to identify words or categories that may usefully be incorporated into the levels of YLE as part of the YLE review. For example, words such as ‘campsite’ and ‘passport’, which appear in the Breakthrough syllabus under the topic of ‘travel’, are currently being considered for inclusion in the YLE wordlists. These wordlists are currently being looked at as part of the Review of the YLE Tests (described in *Research Notes* 15).

Table 2: Overlap of topics for the *Personal* domain

Breakthrough	Starters	Movers
Personal identification	Family, friends & ourselves	Family, friends & ourselves
Education	School & the Classroom	School & the Classroom
Character & personal appearance	The body & the face	The body & the face
Accommodation, rooms	The home	The home
Household articles	–	Containers
Environment	–	World around us
Flora & fauna	Animals	World around us/ animals
Climate & weather	–	Weather
Daily life	Daily life	The home
Entertainment	Leisure-time activities	Leisure-time activities
Sports & physical activities	Sports	Leisure-time activities
Social life	Friends	Leisure

The Young Learner syllabuses and Breakthrough level

In another part of the research, reference is made to the increase of English (and other second language) language learning in the

young learner classroom. This increase suggests that a core language/skills syllabus for young learners is desirable.

Issues raised by the research report include the need for young learner-specific 'can do' statements. Currently Breakthrough lists these for some topics or skills, but not for others, e.g. can do statements exist for 'work & education', but not for 'receptive activities'. The ALTE Young Learners working group is currently working on producing 'can do' statements for young learners. Another issue raised by this research is that 'can do' statements need young learner-specific exemplification. For example, in chapter 6 of Breakthrough, the listening skills that children can be expected to develop are listed, but examples relate to adult contexts such as 'listening to public announcements'.

Conclusion

This comparison between Breakthrough level and Starters and Movers is being used to inform the YLE Review. Revision of the YLE tests will include the selection of new lexical items for inclusion in revised wordlists and consideration of how the lexical content at each level of YLE builds on what is learnt and tested in the previous level.

References and further reading

- Burrow, T and Wilson, J (2004) The YLE Review: findings from a stakeholder survey, *Research Notes* 15, 16–18.
- Council of Europe (2001) *Common European Framework of Reference for Languages: learning, teaching, assessment*, Cambridge: Cambridge University Press.
- Trim, J (in prep) *A1 Breakthrough Level of the Common European Framework of Reference*, commissioned by the Council of Europe.

BAAL/CUP Workshop-Vocabulary Knowledge and Use: Measurements and Applications

Around fifty participants attended the BAAL/Cambridge University Press vocabulary workshop which took place at University of the West of England, Bristol in January 2004. The key themes of the workshop were: comparing different vocabulary measures; describing vocabulary use; teaching and learning vocabulary, and measuring and testing vocabulary.

Roeland Van Hoet (Nijmegen) and Anne Vermeer (Tilburg) considered the reliability and validity of lexical richness measures, using an array of statistical tests and equations. Paul Meara (University of Wales, Swansea) concentrated on one vocabulary measure, the Lexical Frequency Profile. The third keynote, by David Malvern and Brian Richards (both Reading University) presented an

approach using graph theory which seemed the most coherent approach to working out an individual's vocabulary range from one piece of text. The fourth and final keynote was given by Paul Nation (University of Victoria, Wellington, New Zealand) who considered the threats to validity in vocabulary measurement studies. His talk was of particular relevance to language testers. The remaining papers described various vocabulary measures including teaching and learning vocabulary, measuring and testing vocabulary and studies of the vocabulary of particular groups including asylum seekers and the historical increase in vocabulary size.

The resurgence of interest in vocabulary studies is of key interest to Cambridge ESOL.

Making the grade: score gains on the IELTS Writing test

TONY GREEN, RESEARCH AND VALIDATION GROUP

Introduction

IELTS is said to measure 'the language ability of candidates who intend to study or work where English is used as the language of communication' (IELTS 2003:3). As such, it is widely accepted by universities and colleges in the UK and elsewhere as evidence that an applicant has sufficient knowledge of the language to study through the medium of English. It is also widely recognised for purposes of immigration and for professional certification.

Test users are advised to set local score requirements depending on their needs. Educational institutions are advised to take account of factors such as the linguistic demands imposed by different courses (Law, for example, being considered more linguistically demanding than Mathematics) and on the degree of in-session (post-admission) language support they are able to provide to their students. Where otherwise suitably qualified candidates present IELTS scores that do not meet the local criteria, test results are often used to guide decisions about appropriate periods of pre-session (pre-admission) language study.

Incorrect assumptions about the time required to make specified gains could have unfortunate consequences. For example, students, despite studying for the recommended period, might fail to achieve the scores they need to enter the course of their choice. Alternatively, institutions might accept students onto their courses who impose a greater burden on resources than anticipated (see Banerjee 2003 for a discussion of the cost implications for university departments). In this article I will outline research undertaken by the Research and Validation Group and others into the extent and predictability of IELTS Writing score gains: the improvement that learners can expect to make in their IELTS Writing scores over time.

Until 2002 the IELTS partners made the following recommendations regarding score gains, 'It has been shown that individuals can take up to 200 hours to improve by one IELTS band' (IELTS 2002:22). However, this statement was accompanied by caveats and it was stated that score gain was affected by learner characteristics such as age, motivation, first language and educational background.

Advice on Writing score gains

The BALEAP Guidelines on English Language Proficiency Levels for International Students to UK Universities (Bool et al 2003:5) also recommend a period of study time necessary to improve IELTS band scores. In this context the formula is understood to cover longer periods than 200 hours. Thus three months of intensive pre-session English study (approximately 300 hours of classroom

instruction) is said to be sufficient to prepare a student presenting an IELTS score of 5.5 for entry to a 'linguistically demanding' course with an entry requirement of IELTS band 7.

Studies involving IELTS Writing score gains

Notwithstanding the importance of determining appropriate periods of language study, research evidence regarding IELTS score gains is sparse. Coomber (1998) surveyed 100 institutions worldwide and found that none was able to provide any empirical evidence for improvements in IELTS scores among learners on their courses.

Since that time a number of small-scale studies involving IELTS score gains on intensive language programmes (18 hours or more of instruction per week) have been undertaken (Archibald 2002, Brown 1998, Read & Hayes 2003). Of the IELTS test components, Writing score gains, the focus of this article, are the best researched. Although Read & Hayes (2003) included Reading and Listening, Brown (1998) addressed only Writing and Archibald (2002) was restricted to Writing Task 2. All three covered only the Academic module.

Read and Hayes (2003) found that the 17 students participating in their study of IELTS preparation courses in New Zealand made an average improvement on the Reading, Writing and Listening components of 0.36 of a band (from 5.35 to 5.71) following one month of instruction. However, t-tests indicated that, with the exception of Listening scores on one course, these gains were not statistically significant. Brown (1998) compared two 10-week courses: an IELTS preparation course and an English for Academic Purposes (EAP) course with no specific IELTS focus. He found that the nine IELTS preparation students gained by an average of 0.9 of a band on the academic writing module (from 4.3 to 5.2), but that scores for the five students on the EAP course declined by 0.3 of a band over the same period (from 5.3 to 5.0).

In the UK, Archibald (2002) employed a nine-band analytic rating scale, with seven criteria, developed by Henning and Hamp-Lyons (1991) and similar to that used for IELTS. The writing tasks used in the study were closely based on Task 2 of the IELTS Academic Writing module. Participants were found to make an average gain of 1.1 bands (from 4.49 to 5.59) following participation in an eight-week pre-session EAP course. Table 1 summarises the results of the above writing score gain studies.

Although the studies cited above appear to be reasonably consistent with the *two months study: one band gain* formula, there are a number of reasons for caution in interpreting the findings. Brown (1998) and Read and Hayes (2003) acknowledge that the small number of participants involved in their studies limit

Table 1: Summary of results of IELTS studies involving Writing gains

Study	Focus	Number of Participants	Period of Instruction	Average Pre-test Score	Average Post-test Score	Average Score Gain
A: Brown (1998)	Academic	9 (IELTS prep) 5 (EAP)	10 weeks 10 weeks	4.3 5.3	5.2 5.0	0.9 -0.3
B: Archibald (2002)	Academic (Task 2 only)	50 (EAP)	8 weeks	4.49	5.59	1.1
C: Read and Hayes (2003)	Academic (3 skills)	17 (IELTS prep)	4 weeks	5.35	5.71	0.36

the conclusions that can be drawn about score gains. All of the cited studies involved relatively low proficiency learners with average scores on the initial test ranging from 4.3 to 5.35, while many universities require an overall band score of 6.0, 6.5 or higher for admissions purposes. It is not clear whether learners can expect to progress from a score of 5.5 to 6.5 in the same timeframe as those moving from 4.5 to 5.5.

There are a number of further reasons to be cautious in interpreting studies involving score gains. Among other considerations, the following will need to be taken into account: qualities of the test instruments, rater reliability and candidate factors.

1. Qualities of the test instruments

Tests used in IELTS gains studies should involve tasks that accurately represent the content and difficulty of the official IELTS tasks. Cambridge ESOL carry out trialling and statistical equating to ensure that test tasks on each form of the test are comparable. Where unofficial forms are used, evidence should be provided that these accurately reflect the official test. Unless steps are taken to demonstrate that tests given on different occasions are equivalent, any improvements in scores could be attributable to differences in the difficulty of the tests, rather than to improvements in candidate performance. Archibald (2002) used published practice versions of Task 2, allowing 40 minutes for completion, while Read and Hayes (2003) and Brown (1998) used retired versions of the official test. Apparently all three studies used different versions of the test at entry and exit.

2. Rater reliability

Even with appropriate training, the people who score writing tests, the raters, may disagree with each other in the scores they give to the same scripts (or with their own ratings of the same scripts on a different occasion). The effects of disagreements between examiners can be mitigated through the kind of training and sample monitoring carried out by Cambridge ESOL or by double rating. Where techniques of this kind are not used, the changes in scores over time may be attributable to differences in rater behaviour. In cases where raters have also worked as instructors on a course involved in the study, there is a risk that shifting expectations of performance could affect the comparability of the marks given at the beginning and end of the course. Brown (1998) had each script

marked by three IELTS trained raters, Read and Hayes (2003) employed double marking by accredited IELTS examiners while Archibald (2002) used a single marker on each occasion.

3. Candidate factors

A problem besetting all gains studies is the question of response validity. It cannot be assumed that all participants will be equally keen to respond to the test on both occasions. Test takers tend to perform better on a test when the results have meaning for them so that scores for the same group of candidates tend to be lower on practice tests than on tests associated with important decisions. Equally, certain groups of participants may have greater motivation to succeed than others. Brown (1998) acknowledges, for example, that the EAP learners in his study (who would not take the IELTS test for several months) may have been less motivated than their IELTS preparation counterparts who would take the test soon after the end of their course.

Commissioned Research

In response to the growing interest in the question of IELTS score gains and the use being made of the test in determining periods of study, the Research and Validation Group undertook an investigation of the subject in 1999 (Gardiner 1999). Four studies by external researchers have also been commissioned by IELTS partner bodies specifically to cast further light on this question. These include Elder and O'Loughlin (2003), which investigated score gains on components of the Academic Module, Rao, McPherson, Chand and Khan (2003) which explored gains on the General Training (GT) Module, Green and Weir (2002 and 2003) both of which addressed score gains on the Academic Writing test.

Rao et al. (2003:7) used tests of Reading and Writing, 'designed to parallel the IELTS General Training Reading and Writing test modules'. The tests were administered at course entry and exit (although the authors provide no evidence that the tests are, in fact, equivalent in content or difficulty to the official test). Elder and O'Loughlin (2003) entered their participants to official administrations of the IELTS test and Green and Weir (2002) involved a retrospective analysis of results from candidates who had taken the IELTS test on more than one occasion. Green and Weir (2003) used two official forms of the test provided by Cambridge ESOL. The tasks from these two tests were combined to create four test forms, A, B, C and D, used in a crossover design.

Participants and Settings

Rao et al. (2003) included an experimental group of 48 students studying to take the GT module at three institutions in Fiji, and a control group made up of a further 9 individuals, drawn from those registered to take the IELTS test. The experimental group was recruited through the offer of a reduced rate course in IELTS preparation. The courses were delivered in part-time mode over four weekends and totalling 30 hours of classroom instruction.

Elder and O'Loughlin (2003) involved 112 students at four institutions in Australia and New Zealand. Courses ranged from 10 to 12 weeks in length, totalling between 200 and 240 hours of classroom instruction. They included courses in general English and English for academic purposes, but all provided some form of preparation for the Academic module of IELTS.

Gardiner (1999) and Green and Weir (2002) both performed retrospective analyses of results from candidates who had taken IELTS on more than one occasion. Gardiner (1999) looked at both modules and covered the period from 1998 to 1999, while Green and Weir addressed only the Academic Writing module and covered a longer timeframe: January 1998 to June 2001. As no candidate is permitted to retake the IELTS test within a three month period, the interval between tests was at least 12 weeks. Unfortunately, background data collected by the IELTS partners does not include details of whether candidates have spent the period between tests engaged in any form of English study. Hence score gains could not be related to instructional variables.

Green and Weir (2003) was a set of related case studies involving 476 students at 15 institutions. Courses ranged between three and twelve weeks (with periods of between three and ten weeks between tests) and represented variety in the balance of (Academic) IELTS preparation and EAP content.

Scoring

Rao et al. (2003) employed a nine band scale, similar to that used for IELTS, to score their writing tests. However, the number of raters employed to score each script is not stated. The use of more than

one rater is widely recommended to enhance the reliability of measurement. As Elder and O'Loughlin (2003), Gardiner (1999) and Green and Weir (2002) used official IELTS administrations, they were able to rely on the official IELTS scoring procedures. IELTS scripts are not routinely double rated, but in the case of 'jagged score profiles', where a candidate's performance on the writing test is inconsistent with performance on other components, a second rating comes into play.

In Green and Weir (2002) scripts were double rated by accredited IELTS examiners. In addition, rater severity and task difficulty were controlled through application of multi-faceted Rasch analysis using the FACETS computer programme (Linacre 1988). This procedure allows scores to be adjusted to compensate for the relative harshness of examiners and any variation in task difficulty.

Results of Commissioned Studies

A summary of the results of the commissioned studies is set out in Table 2 which displays the mean Writing score gains observed in these studies.

Figure 1 provides a graphical representation of these together with the two studies described above which dealt with Writing. The results displayed in Figure 1 for Brown (1998) include only those learners studying on the IELTS preparation course and not the relatively unsuccessful EAP learners. Results displayed in Figure 1 for Green and Weir (2002) and Gardiner (1999) are limited to those candidates who retook the IELTS test within three to four months of their first attempt.

Figure 1 shows a varied picture of growth and suggests differences between groups of learners. Participants in the Rao et al. group (entrants for the GT module in Fiji) apparently made relatively dramatic progress following a short period of instruction, while learners on the longest courses seemed to make less gain than predicted by the IELTS partner recommendations. Gains appear to be relatively higher for those scoring lowest on the first test.

Gardiner (1999), Green and Weir (2002) and Elder and

Table 2: Mean Writing score gains observed in the IELTS commissioned studies

Study	Focus	Number of Participants	Period of Instruction	Average Pre-test Score	Average Post-test Score	Average Score Gain
D: Rao et al. (2003)	GT	48 (IELTS prep)	3 weeks part-time (30 hours)	5.54 (Task 1) 5.47 (Task 2)	6.02 6.37	0.48 0.90
E: Elder and O'Loughlin (2003)	Academic	112 (various)	10 to 12 weeks intensive	5.08	5.68	0.61
F: Gardiner (1999)	Academic and GT	3,052	Unknown (interval of 3 months or more)	5.12	5.42	0.29
G: Green and Weir (2002)	Academic	15,380	Unknown (interval of 3 months or more)	5.26	5.63	0.37
H(i-iv): Green and Weir (2003)	Academic	476 (IELTS prep and EAP)	3 to 12 weeks intensive (mean 5.5 weeks)	5.27	5.48	0.21

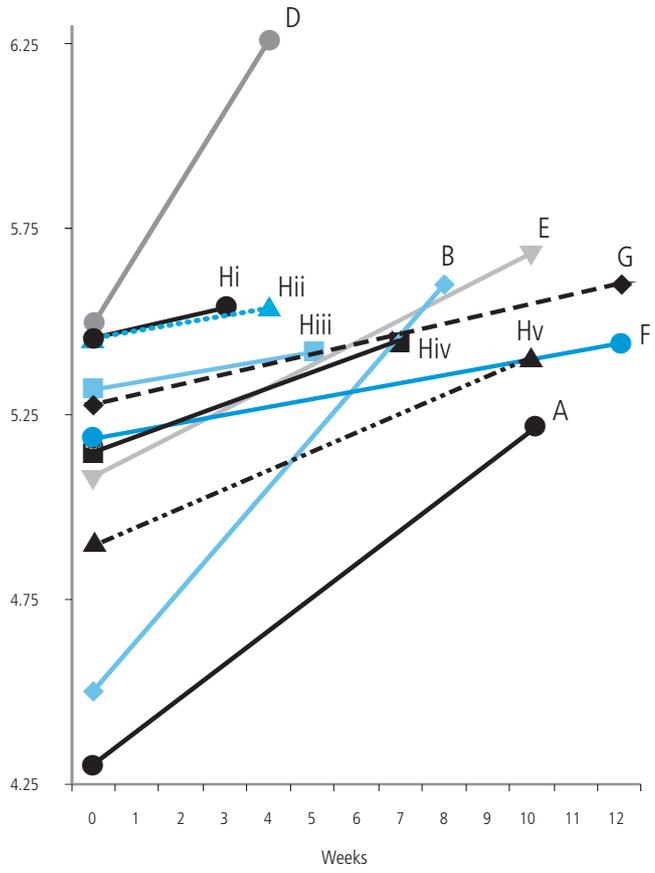


Figure 1: Mean gain in IELTS Writing scores on seven studies

- D: Rao et al
- Hi: Green & Weir 2003*
- ▲ Hii: Green & Weir 2003
- Hiii: Green & Weir 2003
- ◆ G: Green & Weir 2002
- F: Gardiner 1999
- Hiv: Green & Weir 2003
- ▼ E: Elder & O'Loughlin 2003
- ▲ Hv: Green & Weir 2003
- ◆ B: Archibald 2002
- A: Brown 1998

* Green & Weir 2003 numbered i-iv by increasing period between tests

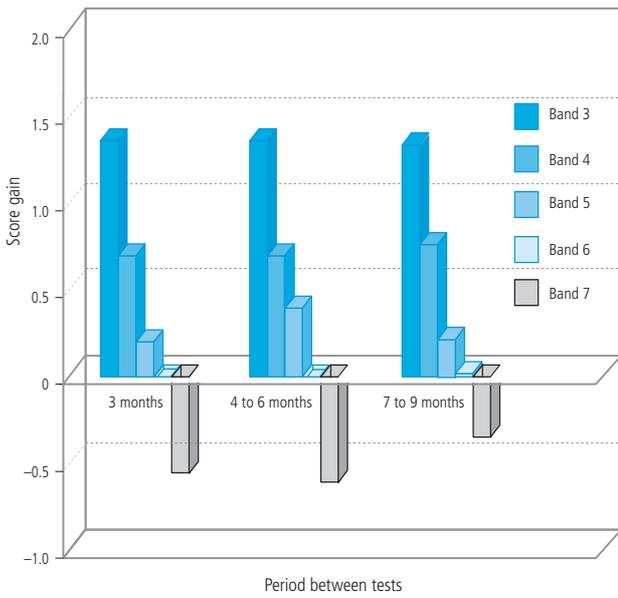


Figure 2: Mean score gains on IELTS Writing component for repeating candidates (GT and Academic) 1998-1999 based on Gardiner (1999)

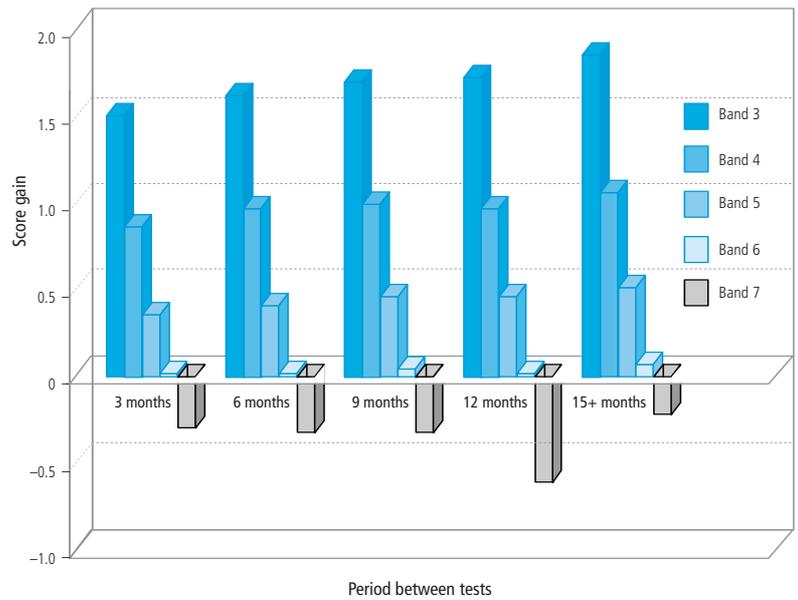


Figure 3: Mean score gains on IELTS Writing component for repeating candidates (Academic) 1998-2001 from Green and Weir (2002)

O'Loughlin (2003) also investigated score gains at different bands. Their results suggest that improvements seen in mean scores do not apply equally at all band levels.

Figures 2 and 3 show that in the Gardiner (1999) and Green and Weir (2002) studies, repeating candidates with Writing scores at band five or below on the initial test tended to improve their results on the second test. Those obtaining a band seven on the first occasion tended to receive a lower band on the second, while those starting on band six tended to remain at the same level. Longer periods between tests do not appear to yield much greater score gains. However, it is not clear from the data how far different lengths of time between tests could be related to periods of instruction or other language learning opportunities.

Green and Weir (2003) and Elder and O'Loughlin (2003), which did investigate language courses, found a limited effect on score gains for hours of instruction. Both studies found that the single strongest predictor of Writing score gain was initial Writing score.

Both Elder and O'Loughlin (2003) and Green and Weir (2003) collected questionnaire data relating to student background. In addition to entry test scores, Elder and O'Loughlin (2003) found that students who believed that Writing was important for their future academic studies made greater gains. Green and Weir (2003) found evidence for the impact of a number of affective factors including self-confidence in writing ability and integration into the host culture. They also found that increasing use of test taking strategies had some effect.

Both studies found evidence for differences between demographic groups. Gains were greatest for those aged in their early twenties and lowest for those from mainland China (in the Elder and O'Loughlin study) or from East Asia (in Green and Weir 2003).

Conclusions

The available evidence suggests that a formula of one band in 200 hours of intensive study cannot be supported. Indeed, period of instruction appears to be rather a poor guide to likely score outcomes.

The IELTS commissioned studies demonstrate the importance of initial status in predicting outcomes. Low level students tend to make gains, while progress for higher level students is either slower, or is not as well captured by the test. It has long been recognised that test scores tend to regress to the mean; that those performing poorly on the first occasion are more likely to improve their scores while those who have done well on the initial test may lose ground. It is also possible that lower proficiency learners benefit more from brief periods of instruction than their higher level counterparts; that band six represents a plateau beyond which learners find it difficult to progress.

Receiving institutions should take note of these findings and reconsider any assumptions they currently make about the relationship between IELTS Writing scores and entry criteria. Two months of intensive English study is unlikely to result in a

score gain on IELTS of one band. On the other hand, opportunities for learning provided by pre-session courses may result in equally worthy outcomes such as familiarity with local study practices and institutional expectations or acclimatisation to the local culture.

References and further reading

- Allwright, J and Banerjee, J (1997) *Investigating the accuracy of admissions criteria: A case study in a British University*, CRILE Occasional Reports 7, Lancaster University.
- Archibald, A (2002) Managing L2 Writing Proficiencies: Areas of change in students' writing over time, *International Journal of English Studies*, 1(2), 153–174.
- Banerjee, J (2003) *Using language proficiency test scores: the admissions process in a British university*, paper presented at the Language Testing Research Colloquium, University of Reading, 22–25 July.
- Bool, H, Dunmore, D and Tonkyn, A (2003) *The BALEAP Guidelines on English Language Proficiency Levels for International Applicants to UK Universities*, London: British Association of Lecturers in English for Academic Purposes.
- Brown, J D H (1998) An investigation into approaches to IELTS preparation, with particular focus on the Academic Writing component of the test, in Wood, S (Ed). *IELTS Research Reports Volume 1*, Sydney: ELICOS/IELTS Australia, 20–37.
- Coomber, J (1998) *Are Test Preparation Programs Really Effective? Evaluating an IELTS Preparation Course*, Unpublished MA dissertation, University of Surrey.
- Criper, C and Davies, A (1988) *ELTS Validation Project Report*, London: The British Council and the University of Cambridge Local Examinations Syndicate.
- Elder, C and O'Loughlin, K (2003) Investigating the relationship between intensive EAP training and band score gain on IELTS, in *IELTS Research Reports Volume 4*, Canberra: IELTS Australia, 207–254.
- Gardiner, K (1999) *Analysis of IELTS Re-take Candidates*, Cambridge: UCLES internal report.
- Green, A and Weir, C (2002) *Monitoring score gain on the IELTS Academic Writing module in EAP programmes of varying duration. Phase 1 report*, Cambridge: UCLES.
- (2003) *Monitoring score gain on the IELTS Academic Writing module in EAP programmes of varying duration. Phase 2 report*, Cambridge: UCLES.
- Hamp-Lyons, L and Henning, G (1991) Communicative writing profiles: An investigation of the transferability of a multiple trait scoring instrument across ESL writing assessment contexts, *Language Learning*, 41(3), 337–373.
- International English Language Testing System (2002, 2003) *The IELTS Handbook*, University of Cambridge Local Examinations Syndicate, The British Council, IELTS Australia.
- Rao, C, McPherson, K, Chand, R and Khan, V (2003) Assessing the Impact of IELTS Preparation Programs on Candidates' Performance on the GT Reading and Writing Test Modules, in *IELTS Research Reports Volume 5*, Canberra: IELTS Australia, 237–262.
- Read, J and Hayes, B (2003) The impact of the IELTS test on preparation for academic study in New Zealand, in *IELTS Research Reports Volume 4*, Canberra: IELTS Australia, 153–206.

Question uptake in the Certificate in Advanced English Writing Paper

FIONA BARKER, RESEARCH AND VALIDATION GROUP
CRIS BETTS, EXAMINATIONS AND ASSESSMENT GROUP

Introduction

The Certificate in Advanced English (CAE) was introduced in December 1991. It is designed to offer a high level qualification in the language to those wishing to use English for professional or study purposes with an emphasis on real-world tasks. The annual candidature for CAE is in excess of 70,000 worldwide and includes candidates from around 90 countries with an average age of 23 (this is lower in certain countries, e.g. Greece).

This article considers the range of task types and topics in the Writing paper and reports the uptake of optional questions in recent administrations.

The CAE Writing paper

CAE candidates are expected to complete two writing tasks of around 250 words in response to input texts and instructions. As in the other Upper Main Suite exams, the CAE Writing paper consists of two Parts. Part 1 is compulsory and candidates must complete one or more tasks in response to a reading input which is usually made up of several short texts. Part 2 involves choosing one of four tasks from a range of writing tasks (letters, articles, messages, reports etc), the last of which is always work-oriented.

The CAE Writing paper is double-marked during on-site marking sessions. Each answer is awarded a mark out of 5 and the two marks are added together to give an overall script mark. All exam scripts are marked by two examiners and a proportion are re-marked by a third examiner where there is discrepancy between the first two ratings. In exceptional cases a script is fourth marked by the Principal Examiner when the third mark is mid-way between the first and second marks, i.e. discrepancy still exists. The number of scripts that are re-marked is very small, amounting to around 5% for third marking and less than 1% for fourth marking. This marking provision ensures that all candidate scripts are marked fairly and reliably. The behaviour of CAE Writing Examiners and the reliability of this paper are not the main focus of this article but will be reported in a future issue of *Research Notes*.

Topics and task types in CAE Writing

Over the last four years (2000–2003) there have been 16 sessions for CAE (two per year for two versions) which have included a total of 64 optional questions, each of which was selected by a minimum of 8% of candidates and a maximum of 64% in any one

administration. The Part 2 questions covered a range of sixteen topics, the most common ones being:

- work
- language learning
- social/national customs
- people

Work, unsurprisingly, was the most common topic (remember that every Question 5 is work related), occurring in 30% of questions, followed by *language learning* and *social/national customs* with 10% each and *people* being the topic of 8% of writing tasks in this period.

A total of thirteen task types were included in the CAE Writing paper. Table 1 shows the percentage of the most frequently occurring task types from 2000–3.

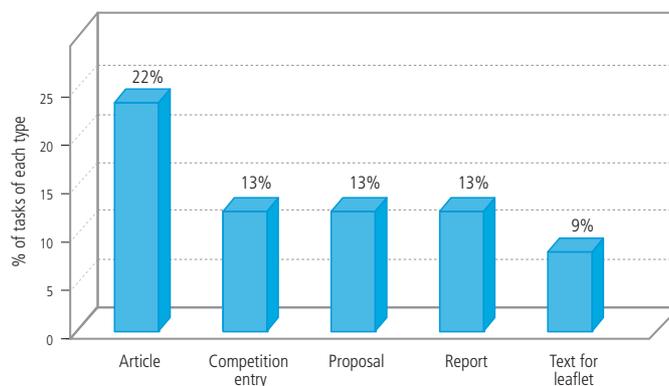


Table 1: Most frequent task types in CAE Paper 2 2000–3

Table 1 shows that the most common format for optional CAE writing tasks has been an *article* (over 20% of questions), followed closely by *competition entries*, *proposals* and *reports* which together account for 40% of writing tasks. Just under 10% of tasks required candidates to write *text for a leaflet*.

From the most common task types and topics in the Writing paper we will now consider what questions candidates choose for Part 2.

Candidate question choice in Part 2

After each exam administration the uptake of optional questions is studied alongside the examiners' ratings of candidate scripts and the range of marks awarded for each question. The uptake of Part 2 questions is useful for determining suitable question formats and

topics for future papers and also for checking the face validity of the tasks that we offer candidates.

In 2000–2003 the most popular questions, each attracting more than half of the candidature in any one session, were a letter on language learning and articles on a range of topics including tourism and technology. Table 2 shows the uptake of questions in the December 2003 administration of CAE.

Table 2: Uptake of questions in CAE Writing December 2003¹

Version	Part	Question	% of candidates
A	1	1	100
		2	44
		3	13
		4	15
		5	28
B	1	1	100
		2	35
		3	28
		4	7
		5	29

¹ Based on a sample of 27,000 candidates.

Table 2 shows us that, in the December 2003 administration, question 2 was the most popular option in both versions, with 44% and 35% of candidates choosing this task. In version A there was a varied spread of question uptake, with question 5 being attempted by 30% of candidates and a smaller though similar number of candidates attempting questions 3 and 4 (13–15%). In version B questions 3 and 5 were chosen by just under 30% of candidates with question 4 being the least popular, being attempted by only 7% of candidates.

These figures are used by Cambridge ESOL to improve the face validity of future question papers for candidates and additionally to determine what factors in a question allow candidates to show their strengths. The interaction of how candidates score on each question is also analysed, as is the rating behaviour of examiners. Similar analyses are done for the productive components of Writing and Speaking for other examinations, although these will differ according to the range of options available to candidates (e.g. Upper Main Suite candidates have an element of choice in both their Writing and Speaking papers) and how each paper is rated and graded.

Conclusion

After every exam administration each Paper is subjected to a range of analyses in order to determine how the question papers, candidates and examiners behaved in that particular session. This article has focussed on the CAE Writing paper and had shown that CAE candidates choose from the whole range of Part 2 questions available and this suggests that at all of these choices are suitable for this candidature.

Whilst Cambridge ESOL does not routinely publish detailed question paper statistics we do provide exam reports for the majority of our examinations which describe how candidates responded to questions, the uptake of optional questions and lists of Dos and Don'ts to help candidates prepare more effectively for each paper. You can also find ideas for skills development in the online CAE Teaching Resource. Here there are plenty of classroom activities to give students practice in developing these skills and to prepare them for the CAE exam. You will also find detailed information about each of the papers and links to useful resources. Please take a look at <http://www.CambridgeESOL.org/teach/cae> To view and download past CAE exam reports, which are freely available and would be of benefit to both teachers and learners, please visit: <http://www.CambridgeESOL.org/support/dloads/index.cfm>

ESOL Staff Seminar programme

ROWENA AKINYEMI & FIONA BARKER, RESEARCH AND VALIDATION GROUP

Cambridge ESOL staff have the opportunity to attend monthly seminars and workshops which form part of the staff development programme. In the past year we enjoyed a varied programme of presentations led by internal and external speakers, ranging from updates on internal policies and practices (e.g. Rasch analysis; the centre network), through exam or market focused talks (e.g. IELTS; developments in the UK) to other ESOL related topics. This article summarises some of the contributions to the past year's programme made by external speakers.

Vocabulary issues in language testing

Norbert Schmitt (University of Nottingham) led a seminar on vocabulary in which he raised a number of issues that impact on the rationale and practicalities of test development, such as the links between vocabulary size and language use and how test items address aspects of word knowledge.

Of particular interest was the discussion of how much vocabulary a person needs to operate effectively in English.

Australian research has found that 2,000 word families were needed to know 99% of spoken discourse (Schonell et al. 1956). In comparison, research based on the British CANCODE corpus found that 3,000 word families is a more appropriate goal (Adolphs & Schmitt 2003). For reading, where text is denser and a knowledge of lower frequency vocabulary is more important, research has indicated that as many as 5,000 word families are needed, with which 5% of words in a passage would be unknown. For studying in an English medium environment, more than 10,000 word families are needed (this includes technical vocabulary).

The notion of a word knowledge framework was discussed which includes many facets of word knowledge such as different forms and inflections, grammatical patterns and register constraints (Nation 1990). When acquiring language, some aspects of vocabulary knowledge are learnt quicker than others, for example a word's spelling is learnt once early on whereas meanings are learnt over time. Language testers therefore need to consider what is being measured according to this framework.

The question of where and how Cambridge ESOL tests vocabulary was then considered. The CPE Use of English paper, for example, requires collocational knowledge as it asks candidates to identify one word that fits three different contexts. Cambridge ESOL is continually reviewing how it tests vocabulary and one way of doing this is explore corpus data for evidence of how learners and native speakers use vocabulary. Some ongoing research into productive vocabulary has established lists of words from candidate writing taken from Main Suite and BEC exams (see David Horner & Peter Strutt's article in *Research Notes* 15). In future lexical sequences could be identified and used for examiner training. In relation to this, one observer commented that examiners may be tempted to pay attention to "sparkly words" when rating writing or speaking performances. Whilst this is not an issue for objective tests of vocabulary where there is only one correct answer it is important for rater training (see Stuart Shaw's article on IELTS).

We are also involved in specifying the vocabulary that question writers can use which takes into account frequency, word senses and collocational information.

Vocabulary is clearly an important aspect of the construct of language proficiency that lies at the heart of what Cambridge ESOL tests and is particularly relevant to the productive skills of speaking and writing. This seminar was a timely reminder of this and the subsequent discussion touched on many facets of learning and testing vocabulary including the debate of whether items in a listening test have to be spelt correctly and how best to test productive and receptive vocabulary knowledge.

Understanding Classroom-based assessment: new territories for language testers?

Pauline Rea-Dickins (University of Bristol) led a seminar which drew on a recent research project called Classroom Assessment of

English as an Additional Language: Key Stage 1 Contexts (ESRC Research grant R000238196). Pauline presented findings that highlighted different facets of classroom assessment including teachers' understanding of assessment; teacher decision-making; standards and criteria that inform teacher assessment; teacher feedback and learner strategies during classroom assessment.

The research studied the formative assessment of several learners by teachers and language support assistants (LSAs) across the early primary curriculum over a year.

A broad socio-cultural approach was adopted that included direct and recorded observation, analysis of children's work and interviews with teachers and learners. The findings revealed many examples of teachers supporting learners although the large number of assessment procedures seemed a burden for teachers, LSAs and pupils alike.

Pauline presented a conceptualisation of classroom-based language assessment for English as an Additional Language (EAL) which hopefully will influence policy in this important area. She concluded that links remain to be made between formative assessment and language acquisition. One area of particular interest was learner involvement in the formative assessment process and one key point that arose was that classroom assessment should be formative both for the teacher and for the learner.

This session provided informative insights into what happens in primary classrooms in the EAL context, an area which many Cambridge ESOL staff are not directly connected with. Because we concentrate on language testing, it is important to keep in touch with what is happening in classrooms worldwide which clearly includes UK classrooms. In the UK context the adult ESOL field is growing and this presentation helped to raise issues that relate to this area. This seminar and subsequent article (in issue 14 of *Research Notes*) served to raise our awareness of some of the wider language awareness issues that underpin assessment and which should continue to inform our language testing work.

Tasks and the assessment of spoken language

Peter Skehan (Kings College, London) led a session in which he described research into oral task design and explored the relevance for language testing. He reviewed some of the key literature on first and second language acquisition and presented the results of recent research which shows that certain task features have an impact on performance, all of which imply that oral assessment procedures should take account of task-based research.

When testing spoken language, three facets are being tested: fluency (focusing on meaning), accuracy and complexity (both of which focus on form). There is always a trade-off between the three areas with accuracy and fluency competing, and both fighting with complexity. Each facet of language is linked to specific goals in task-based instruction: the goal of real-time accessibility is linked to fluency; control and conservatism are linked to accuracy and risk-taking and extension are linked to complexity.

The design of oral tasks influences different types of performance, as complexity of language, communicative conditions and cognitive complexity play a part in what sort of oral performance is produced. All oral tasks should have certain characteristics: they should be meaning driven, outcome oriented, problem solving and 'real-world'. The notion of task difficulty relates to a number of factors including abstract or unfamiliar information and complex retrieval whereas familiar tasks achieve greater accuracy. Also, oral tasks involving more than one candidate tend to increase the accuracy of both candidates as they scaffold each other.

Several studies on planning strategies were reported which found that students who are given time to plan oral activities may provide more complex language and increased fluency although their language may also be less accurate. The amount of planning time also has an effect as it interacts with task structure, with certain tasks benefiting from extended planning time. Interestingly, it was reported that a lower ability group who were given time to plan an oral activity were more fluent than a higher ability group who did not plan.

Scoring spoken tasks is the result of many interactions between the task characteristics and conditions in which the task is undertaken. We all attempted to rank in terms of difficulty a number of narrative tasks whilst also considering their effect on accuracy, fluency and complexity. The easiest story to narrate had a clear time-line and structure.

The audience discussed task difficulty versus performance quality and how the transparency of a storyline may be less to do with complexity and more to do with foreground/background elements, task selection and interpretation. Differences were also proposed between performance testing and objective testing tasks. The importance of the front-end of speaking test tasks was mentioned (i.e. training item-writers and oral examiners) as were predictions of success and outcomes. Proficiency level per se has not been investigated in the task literature and Peter proposed that this is a suitable topic for future research.

The conclusions of this session were that tasks have both pedagogical and assessment implications, the latter being most relevant for the audience. Whilst there are different implications for accuracy, complexity and fluency, complexity was the defining measure in terms of assessing oral tasks of the type presented and discussed during this session.

Englishes and Identities

John E. Joseph (The University of Edinburgh) and Elizabeth J. Erling (Freie Universität Berlin) delivered a seminar on the subject of 'Englishes and Identities'. John is particularly interested in how people conceive of and use language in their everyday lives, how these conceptions and uses align (or fail to align) with those developed by linguists and philosophers over the centuries, and how the gap between the two can be narrowed, so as to put theories of language into a more useful dialogue with linguistic

cultural practices such as language teaching, translation and rhetoric.

The seminar built around the research John Joseph has been doing for several years on recasting notions of 'standard language' within the discourse of language and identity, and on the work of Elizabeth J. Erling who has been researching the attitudes and motivations of university students of English in Berlin.

John began by surveying how the role of language is differently conceived in various scholarly approaches to nationalism. He gave particular attention to the thesis of Hobsbawm (1991), according to whom our general conceptions of both nation and language are historical products dating back only to the 1880s, closely bound up with social class and class mobility and with Victorian notions of respectability. Linguistic correctness can be understood as a manifestation of 'banal nationalism' (Billig 1995), tacit enactments of national identity through performance and enforcement of norms of 'respectable' language. The 1880s were also the high point of the British Empire, and John looked at how imperial language policies changed around this time, paying particular attention to the case of Hong Kong. The role of English there, together with the ways it has been conceived, was traced to the present day. In Hong Kong, as in most of the developed regions of East and Southeast Asia, English plays a highly-valued identity function for those who speak it. The status and characteristics of 'Hong Kong English' (HKE) was considered through an examination (with audience participation) of samples from scripts produced by students at the University of Hong Kong. John proposed a three-way distinction for dealing with discrepancies from Standard British English based upon their capacity for functioning as markers of identity.

Elizabeth's presentation turned the focus to continental Europe, where the identity functions of English overlap somewhat with those of East Asia, but not entirely. She considered the concept of 'Global English' and then presented the results of her research into what the university students themselves think about the language, its present and future role in their lives, and the precise form of the language they want to produce. English plays an important role in the lives of these students, both inside and outside the classroom: they use the language daily in varied contexts and they are regularly exposed to the diversity of English through internet technology, the international atmosphere of Berlin and regular travels and stays abroad. However, even though these students are enrolled in philology-based English university courses, the type of English they are interested in acquiring is not necessarily based on native norms. One third of the language students see English as a global lingua franca and are not particularly interested in learning about English-speaking cultures nor in identifying with an English-speaking country. These attitudes represent changes in perceptions of English as a result of globalization and suggest that these developments should be addressed by language educators.

The seminar ended with a revisiting of the general issues of language and identity in the light of the specific examples subsequently presented; and then John commented on the role of

language testing as having an impact on the issues which were discussed.

A lively discussion followed the presentations. Questions raised included whether the parameters by which English is assessed by Cambridge ESOL examinations should be widened; how the accommodation of candidates in the collaborative task of the Speaking Tests should be assessed, and whether students' concept of global English enables them to modify their local English. This discussion was felt to be of such interest that a further session was arranged for Cambridge ESOL staff to enable these issues to be discussed in greater depth. This discussion considered the thesis of Hobsbawm (that linguistic correctness can be understood as a manifestation of 'banal nationalism', tacit enactments of national identity through performance and enforcement of norms of 'respectable' language) and went on to discuss whether the construct of Global English could be realised in a way that would be useful and practicable for us for examination purposes.

Insights into Chinese learners of English

Helen Spencer-Oatey (eChina Learning Programme Manager, UK eUniversities Worldwide) presented a seminar entitled ' "... travellers on the world's horizons": insights into Chinese learners of English'. Until very recently, Helen was Principal Lecturer in Linguistics at the University of Luton, responsible (*inter alia*) for the design and management of its MA course in Intercultural Communication. Much of her research and teaching has revolved around Chinese learners, having worked in China for over 10 years (7 of which were in Shanghai) both as an EFL teacher and as a teacher trainer. Her rich cultural experiences in China stimulated her interests in cross-cultural and intercultural issues, and she has published chapters and articles on cross-cultural pragmatics, intercultural discourse, and cross-cultural psychology.

The aim of the seminar was to explore how cultural factors impact on the learning of English in China. Helen outlined the Chinese educational system at primary, secondary and tertiary levels, describing key and non-key schools and universities, and

then going on to examine the teaching and testing of English at each of these levels. Helen described the components of the College English Test, and used extracts from Chinese textbooks, sample test papers and video recordings of authentic English language classes to illustrate Chinese methods of teaching and testing, particularly at university level.

In the second part of the seminar, Chinese learning strategies were considered, especially the use of memorisation and repetition. Helen discussed differing cultural views on memorisation: the Western view that memorisation is negative, hindering productive thinking and problem solving; and the Chinese view that memorisation is the first way to learn and that repetition is linked to creativity.

Questions were raised by ESOL staff at different points during the seminar, resulting in interesting discussions on, among other topics, the role of the teacher in China, which encompasses a broad set of responsibilities, and the attitudes towards formality in Chinese culture. At the end of the seminar we all felt that Helen had taken us on an interesting journey towards the horizon of learners of English in China.

These seminars reflect the wide range of topics which Cambridge ESOL staff are interested in. The 2004 programme promises a selection of topics from language testing and beyond presented by a range of internal and external speakers, reviews of which will appear in future *Research Notes*.

References and further reading

- Adolphs, S and Schmitt, N (2003) Lexical Coverage of Spoken Discourse, *Applied Linguistics*, 24/4, 425–438.
- Billing, M (1995) *Banal nationalism*, London: Sage.
- Hobsbawm, E J (1991) *Nations and nationalism since 1780: Programme, myth, reality*, Cambridge: Cambridge University Press.
- Nation, I S P (1990) *Teaching and learning vocabulary*, Boston: Heinle & Heinle.
- Schonell, F, Meddleton, I, Shaw, B, Routh, M, Popham, D, Gill, G, Mackrell, G and Stephens, C (1956) *A Study of the Oral Vocabulary of Adults*, Brisbane and London: University of Queensland Press/ University of London Press.

Conference Reports

Cambridge ESOL staff attend a wide range of conferences and seminars each year. The following reports describe two BALEAP events attended by Cambridge ESOL staff.

BALEAP Conference 2003: Developing Academic Literacy

The BALEAP Conference, held at the University of Southampton in April 2003 was a well-attended and stimulating event.

160 delegates from universities in UK, Canada, Australia, Turkey, Hungary and Greece took part and there were over 50 sessions. The theme of the conference was 'Developing Academic Literacy' and the speakers approached this from a variety of different angles. The issue of plagiarism was addressed in several sessions. Nigel Harwood analysed the use of citation from both an applied linguistics and information science/sociology perspective and showed how students overuse certain kinds of citation. He argued that for students to become academically literate, they need to

develop a deeper understanding of how and why authors cite. Joan McCormack and Hania Salter-Dvorak discussed practical teaching activities aimed at sensitising students to the concept of plagiarism and showing them ways of avoiding it. Hania Salter-Dvorak suggested that the reason for the prevalence of plagiarism in academic writing may be due to a lack of understanding of this genre and of the concepts of authorship, copyright and intertextuality.

Related to the issue of plagiarism was Fiona Cotton's session on critical thinking. Feedback from academics reports the lack of a critical voice in students' assignments. Fiona's study compared the use of evaluative language in 'native' and 'non-native' speaker graduate assignments. It was clear that non-native speakers used fewer instances of evaluative language. Interestingly, the lack of a critical voice in some ways seems separate from language proficiency. Students at IELTS Band 7 or 8 often do not show a 'critical' voice in their writing.

Other sessions focused on the importance of speaking in the academic curriculum (Barbara Hughes); the development of an academic writing course at UCL (Simon Williams and Steve Marshall) and the processes and principles of an integrated reading and writing course for pre-sessional students of IELTS level 5 and upwards at Reading University (Anne Pallant and John Slaght). Dilys Thorpe described a British Council funded project exploring how candidates responded to IELTS Academic Writing Task 2 (a 250 word essay responding to a point of view, argument or problem). Using a discourse analytic approach she outlined the characteristics of scripts at three levels on the nine-band IELTS scale (band four, band six and band eight) and considered how far these responses reflected academic literacy practices expected of university work.

BALEAP Professional Issues Meeting: EAP Online

A BALEAP (*British Association of Lecturers in English for Academic Purposes*) Professional Issues Meeting was held at the University of Newcastle on Saturday 14th June 2003. The one-day event, one of a sequence of PIMs addressing technology issues, focused on the topic of EAP On-line, or the use of internet technologies in the delivery of English for Academic Purposes courses.

Alex Ding from the University of Nottingham opened the meeting with a talk entitled 'Issues in creating effective on-line collaborative learning environments for EAP students'. Alex described some of the challenges involved in fostering collaborative learning in the on-line environment and actions taken to address these. The issues raised by Alex recurred throughout the meeting. These included the need to train learners to make effective use of the online environment, translating traditional materials to the virtual classroom, adapting course design to the needs of teachers and learners and coping with limited resources and technical constraints. John Wrigglesworth described the integration of a Virtual Learning Environment (VLE)

into a fifteen-week foundation course, and the commitment of lecturer time that this demands. Sandy MacIntosh (University of Alberta) described the development of online courses at his institution and suggested that an online course could never be regarded as a finished product, arguing from experience for constant management and revision. Harriet Edwards (Royal College of Art) described the personal investment and the learning processes involved for the tutor in developing web-based support material for Art and Design students. Each of these papers demonstrated the importance of commitment on the part of individuals to the success of online learning tools.

Andy Gillett (University of Hertfordshire) spoke of university-wide initiatives putting pressure on departments to employ VLEs without a principled basis for integrating these into the curriculum. Andy suggested a model of teaching and learning processes in Higher Education that could be used to guide the integration of VLEs. The model would allow course developers to consider how teachers, learners and administrators might use a VLE to enhance each phase of the teaching/learning cycle. In a second session, Andy introduced a number of readily available low-cost software tools that could be used to develop interactive computer-based activities and led a discussion on how these could be deployed most effectively in EAP contexts.

In a different vein, Eddy Moran described an experiment into how students made use of a vocabulary learning programme. The programme introduced learners to a series of individual words. For each word, learners were given a number of options including seeing a dictionary definition, attempting to guess the meaning in context, or selecting the correct meaning from a number of choices. The level of prior knowledge of the word seemed to play a central role in how the learners used the programme. Students explored the most options when they had some prior knowledge of a word. When students either had no prior knowledge, or when they reported familiarity with the word, they tended to move directly to the dictionary definition. Eddy also found significant correlations between belief in the relative importance of accuracy or fluency and the amount of effort learners invested in learning.

The event closed with a visit to the University of Newcastle's impressive Open Access Centre where delegates had the opportunity to try out some of the resources they had heard about and to experience the student's eye view of EAP online.

References and further reading

- For information on past and future BALEAP Professional Issues Meetings: <http://pims.baleap.org.uk/>
- Andy Gillett's online resources for EAP: <http://www.uefap.co.uk/>
- Chappelle, C A (2001) *Computer Applications in Second Language Acquisition: Foundations for teaching, testing and research*, Cambridge: Cambridge University Press.
- Salmon, G (2000) *E-Moderating: The key to teaching and learning online*, London: Kogan Page.

Other News

New research area of Cambridge ESOL website

There is now a specific Research area on the Cambridge ESOL website which profiles the work of the Research and Validation Group. Cambridge ESOL has the largest dedicated research team of any UK-based provider of English language assessment. The work of this team helps ensure that our exams are fair to test takers whatever their backgrounds and provide an accurate measure of their true abilities. Research and Validation staff have a coordinating role within the organisation for the ongoing research and validation programme for all the ESOL exams.

The Research area includes an overview of our research and validation activities, historical background, key staff biographies, together with information on the Studies in Language Testing series and *Research Notes*. More detailed information on specific projects, publications and presentations will be added in the near future.

In addition, articles written by Cambridge ESOL staff regularly appear in a range of publications related to language testing and teaching. The following have been published recently:

- Fiona Barker, Using corpora in Language Testing, *Modern English Teacher* April 2004
- Andrew Nye, Preparing for ICELT in Mexico, *EL Gazette* April 2004
- Nick Saville and Roger Hawkey, The IELTS Impact Study: Investigating washback on teaching materials, chapter in Cheng, L, Watanabe, Y & Curtis, A (2004) *Washback in language testing: Research contexts and methods*.
- Lynda Taylor, Testing Times: Research Directions and Issues for Cambridge ESOL Examinations, *Current Issues in English Language Testing Research, Tesol Quarterly* Spring 2004

Please visit the website at: www.CambridgeESOL.org/research

New staff in Research and Validation

Several new staff have joined the Research and Validation Group recently. In February we welcomed two new Validation Officers: Peter Hardcastle whose role is to support the examinations of certain ALTE partners and Dr Evelina Galaczi who has responsibilities for Upper Main Suite, Teacher Awards and the new ESOL Skills for Life suite. We also have a new Validation Assistant, Lucy Chambers, who replaces Jenny Jones in organising anchor testing for many of our exams.

Teaching resources for PET, CAE and BEC Vantage

Teachers preparing students for the updated Preliminary English Test, Certificate in Advanced English and BEC Vantage can now find a wealth of information about this exam on the Cambridge ESOL Teaching Resources website. Each Teaching Resource covers:

About the Paper

- Sample tasks from past papers
- Activities to understand the tasks
- Tips for teachers and students
- DOs and DON'Ts

Classroom Activities

- Ready-made activities to use with your class
- Skills development
- Exam practice with tasks from real past papers

Other Information

- Links to other Cambridge ESOL resources
- Handbooks, past papers, videos and CDs

Frequently Asked Questions (FAQs)

- How many answers must the candidate produce?
- How many marks is each question worth?
- Where do students write their answers?
- How is the Writing paper marked?

For more information on Cambridge ESOL's Teaching Resources, please visit www.CambridgeESOL.org/teach

Research Notes offprints

Over 100 *Research Notes* articles are now available to view or download from the Cambridge ESOL website. The offprints from issues 7–15 are listed below; those from issues 1–6 were listed in issue 13.

Issue/Focus	Date	Title	Author/s
ISSUE 7 Young Learners	Feb-02	Developing English language tests for young learners Candidate performance in the Young Learners English Tests in 2000 Research projects relating to YLE Speaking Tests Striving for fairness - the ALTE Code of Practice and quality management systems Investigating variability in a test of second language writing ability Review of KET & PET Examinations Report on the BAAL/CUP Seminar 'Young Language Learners: Towards a Research Agenda' Announcement of the winner of the IELTS MA Dissertation Award 2001	Lynda Taylor & Nick Saville Helen Marshall & Mike Gutteridge Fiona Ball & Juliet Wilson Nick Saville Barry O'Sullivan Nigel Pike & Liz Gallivan Fiona Ball
ISSUE 8 Business English	May-02	Some theoretical perspectives on testing language for business Revising the Business English Certificates (BEC) speaking tests Revising the BULATS Standard Test Developing wordlists for BEC The effect of training and standardisation on rater judgement and inter-rater reliability Investigating gender differences in young learner performance Investigating test conditions for listening and speaking	Barry O'Sullivan David Booth Ed Hackett Fiona Ball Stuart Shaw
ISSUE 9 CELS	Aug-02	Plurilingualism, partial competence and the CELS suite Background to CELS: the communicative construct and the precursor exams The test development process for CELS CELS Writing: test development and validation activity CELS Speaking: test development and validation activity IELTS Writing: revising assessment criteria and scales (Phase 1) Investigating the CPE word formation cloze task Reviewing the retraining of BEC Oral Examiners	Lynda Taylor Roger Hawkey Nick Saville Stuart Shaw & Sharon Jordan Lynda Taylor & Stuart Shaw Stuart Shaw
ISSUE 10 Main Suite	Nov-02	Innovation and continuity: CPE - past and present Redeveloping Part 1 of the CPE Listening paper Update on changes to the KET/PET Writing papers from 2004 IELTS Writing: revising assessment criteria and scales (Phase 2) Linking YLE levels into a single framework Investigating the YLE story-telling task Assessing learners' English: but whose/which English(es)? Exploring issues in the assessment of pen-and-paper/computer-based IELTS Writing Lexicom@ITRI: a Lexicography Course Monitoring oral examiner performance in FCE Monitoring IELTS test performance in 2001 Monitoring speaking test materials for Young Learners Tests	Cyril Weir Rod Boroughs Liz Gallivan Stuart D Shaw Neil Jones Fiona Ball Lynda Taylor
ISSUE 11 Special Circumstances	Feb-03	Responding to diversity: providing tests for language learners with disabilities Producing Modified Versions of Cambridge ESOL Examinations Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts Task difficulty in the assessment of writing: Comparing performance across three levels of CELS Insights into the FCE Speaking Test Converting an Observation Checklist for use with the IELTS Speaking Test	Lynda Taylor & Mike Gutteridge Ruth Shuter Stuart Shaw Neil Jones & Stuart Shaw Yang Lu Lindsay Brooks
ISSUE 12 Technology	May-03	The Role of Technology in Language Testing Electronic Script Management: towards on-screen assessment of scanned paper scripts A quick review of the English Quick Placement Test Recent Developments in Learner Corpora Assistive Technology for Candidates with Special Needs Feedback on CPE re-training	Neil Jones Stuart Shaw Ardeshir Geranpayeh Fiona Barker Mike Gutteridge Chris Hubbard
ISSUE 13 Speaking	Aug-03	The Cambridge approach to speaking assessment What constitutes a basic spoken vocabulary? The development of a set of assessment criteria for Speaking Tests CELS Speaking Assessment: towards an understanding of oral examiner and test-taker behaviour Evaluating the success of the revised BEC (Business English Certificate) Speaking Tests	Lynda Taylor Michael McCarthy & Ronald Carter Angela French Stuart Shaw David Booth
ISSUE 14 Teaching Awards	Nov-03	Cambridge ESOL Teaching Awards: current perspectives, future trends The Distance DELTA DELTA by Distance Learning Diaries, theory, practice and assessment: the teacher educator as reflective practitioner Language Awareness and Assessment In-service language teaching in Brazil using ICELT Teacher Support Interaction in a paired speaking test: the case of the First Certificate in English	Monica Poulter David Albery Dave Russell Craig Thaine Pauline Rea-Dickins Lizika Goldchleger Jill Grimshaw Evelina D. Galaczi
ISSUE 15 Specialised Tests	Feb-04	Issues of test comparability Analysing domain-specific lexical categories: evidence from the BEC written corpus IELTS Writing: revising assessment criteria and scales (concluding Phase 2) An IELTS Impact Study: implementation and some early findings The YLE Review: findings from a stakeholder survey Creating a virtual community of assessment practice: towards 'on-line' examiner reliability Reliability in First Certificate in English objective papers Announcement of the winner of the IELTS Master's Award 2003	Lynda Taylor David Horner & Peter Strutt Stuart Shaw Roger Hawkey Trish Burrow & Juliet Wilson Stuart Shaw Ardeshir Geranpayeh

IELTS joint-funded research 2004/5 (Round 10): call for proposals

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

As part of their ongoing commitment to IELTS-related validation and research, IELTS Australia and the British Council are once again making available funding for research projects in 2004/5. For several years now the two partners have issued a joint call for research proposals that reflect current concerns and issues relating to the IELTS test in the international context. A full list of funded research studies conducted between 1995 and 2001 (Rounds 1–7) appeared in *Research Notes* 8 (May 2002) and later rounds are listed on the IELTS website. Such research makes an important contribution to the monitoring and test development process for IELTS; it also helps IELTS stakeholders (e.g. English language professionals and teachers) to develop a greater understanding of the test.

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

What areas of interest have been identified?

The IELTS Research Committee have identified the following as areas of interest for external research purposes:

- studies investigating the use of IELTS and IELTS scores in local contexts (i.e. studies with different cohorts in terms of age, L1, nationality, use of IELTS in professions, use of IELTS in new regions e.g. USA)
- studies investigating IELTS and its impact on the teaching/learning environment (e.g. classroom observation studies)
- studies investigating IELTS and its impact on teaching/learning materials (e.g. analysis of coursebooks)
- studies investigating perceptions of and attitudes to IELTS among key stakeholder groups (teachers, administrators, test-takers and other test users)
- small-scale, in-depth case studies focusing on individuals or small groups taking/using IELTS
- other issues of current interest in relation to IELTS.

A list of completed funded projects can be found on the IELTS website – www.ielts.org

Is access to IELTS test materials or score data possible?

Access to IELTS test materials or score data is not normally possible for a variety of reasons, e.g. test security, data confidentiality. However, a limited amount of retired material (e.g. writing test prompts) may be made available for research purposes, and IELTS Specimen Materials and published practice tests can often be used as research tasks. Cambridge ESOL may be able to supply writing scripts and speaking test recordings for the purposes of analysis, and a set of instruments and procedures for investigating the impact of IELTS on materials and on the teaching/learning context has also been developed in recent years; these may be made available for use by researchers following consultation with Cambridge ESOL (subject to an appropriate research agreement).

Who can submit proposals?

As part of the IELTS policy of stimulating test-related research among its stakeholders, it is hoped that many of the research proposals submitted this year will come from researchers and organisations who have a direct and ongoing connection with IELTS, e.g. consultants, examiners, IELTS Administration Centres and centres which have assisted in trialling IELTS. There is, however, no objection to proposals being submitted by other groups/centres/individuals.

What is the level and duration of funding available?

The maximum amount of funding which will be made available for any one proposal is £13,000/AUS\$30,000. The research study will need to be completed and a full report submitted by the end of December 2005.

What is the procedure for submitting proposals?

Application forms and guidelines for submission (together with terms and conditions) are available from the British Council and IELTS Australia – see below for contact details. Proposals for funding should take the form of a typed/word-processed document of no more than 10 pages, and be accompanied by the completed application forms.

Who will evaluate the proposals?

All research proposals will be evaluated by the IELTS Joint Research Committee comprising representatives of the three IELTS partners as well as other academic experts in the field of applied linguistics and language testing.

What criteria will be used to evaluate proposals?

The following factors will be taken into consideration when evaluating proposals:

- Relevance and benefit of outcomes to IELTS
- Clarity and coherence of proposal's rationale, objectives and methodology
- Feasibility of outcomes, timelines and budget (including ability to keep to deadlines)
- Qualifications and experience of proposed project staff
- Potential of the project to be reported in a form which would be both useful to IELTS and of interest to an international audience.

What is the time scale for the submission and evaluation of proposals?

The following time scale will apply for Round 10:

May 2004	Call for proposals
31 July 2004	Deadline for submission of proposals
Aug/September 2004	Preliminary review of proposals by IELTS partners
October/November 2004	Meeting of IELTS Research Committee to evaluate and select successful proposals
December 2004	Applicants notified of the IELTS Research Committee's decision

Application forms and submission guidelines are available from:

Ms Sasha Hampson Program Manager Testing Services IELTS Australia IDP Education Australia GPO Box 2006 Canberra ACT 2601 Australia	Ms Sujata Saikia IELTS Business Development Manager British Council 10 Spring Gardens London SW1A 2BN United Kingdom
Tel: 61 6 285 8222 Fax: 61 6 285 3233 E-mail: Sasha.Hampson@idp.com	Tel: +44 (0)20 7389 4870 Fax: +44 (0)20 7389 4140 E-mail: sujata.saikia@britishcouncil.org

Award presentations at LTRC 2004

The following presentations were made at the 26th Language Testing Research Colloquium held in Temecula, California in March 2004.

IELTS Master's Award 2003

In *Research Notes 15* we announced the winner of the IELTS Master's Award 2003. Eunice Eunhee Jang was presented with her certificate and a cheque for \$1700 at LTRC. The photograph shows Eunice receiving her award from Beryl Meiron (Director of IELTS Inc in the USA and North America) on behalf of the three IELTS partners.



Details of the application process for the IELTS Master's Award 2004 can be found on the IELTS website (www.ielts.org) or in *Research Notes 15*.

UCLES-ILTA Lifetime Achievement Award 2004

Professor Lyle Bachman, Professor and Chair of Applied Linguistics and TESL at the University of California, Los Angeles, was presented with the UCLES-ILTA Lifetime Achievement Award at LTRC. Nick Saville read a short citation before presenting the award in which he described Lyle's long relationship with Cambridge ESOL and contributions to the field.

Cambridge ESOL and Lyle go back to around 1987 when Mike Milanovic and Nick Saville were involved in data collection for the Cambridge-TOEFL Comparability Study which was subsequently written up by Lyle and the other researchers and became the first

volume in the Cambridge Studies in Language Testing series. The Comparability study led to several collaborative projects in the early 1990s with Lyle and his colleagues and students at UCLA. Lyle also ran summer courses in Cambridge in 1991 and 1992 for new research staff, on one occasion with Fred Davidson. It is thanks to Lyle we have developed long lasting relationships with some of the best qualified and productive academics in language testing.

There are three key aspects of those early days in the 1990s which sum up Lyle's contribution and which make this award so richly deserved:

- His rigorous and innovative research in the field of language testing.
- His skill and commitment as a teacher and tutor, including his extensive writing and pedagogic materials.
- His inspirational leadership and guidance for a whole generation of language testers – not just in the USA, but around the world. Several of us in Cambridge are happy to count ourselves among that group.

Nick ended by wishing Lyle many more achievements in the field of language testing.



Nick Saville (Cambridge ESOL) Lyle Bachman (UCLES-ILTA Lifetime Achievement Award holder), Anthony John Kunnan (ILTA President 2004)