ResearchNotes

Contents

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

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

Editorial Notes

Welcome to Issue 11 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

2003 is the *European Year of People with Disabilities (EYPD)* so it is entirely fitting that our first issue for this year should contain a special focus on the testing provisions we offer to candidates with special needs. Lynda Taylor and Mike Gutteridge describe the context and work of the Special Circumstances Unit at Cambridge ESOL; this is the unit charged with meeting the needs of candidates who require *special arrangements* (i.e. modified tests for reasons of temporary/permanent disability) or *special consideration* (i.e. special consideration in the light of extenuating circumstances surrounding the testing event). In a follow-up article, Ruth Shuter explains in more detail the complex and lengthy process of producing modified versions of the Cambridge ESOL examinations which meet the same rigorous standards as the standard test papers.

Continuing our strand on writing assessment, Stuart Shaw reports on a recent study to investigate the influence of handwriting on the rating of second language writing performance. The findings of this and similar studies have important implications in the context of computer-based testing for candidates who key rather than handwrite their responses to writing tasks. Also on the theme of writing assessment, Neil Jones and Stuart Shaw explore the issue of task difficulty, especially as it operates across three distinct but overlapping proficiency levels. They highlight the importance of using a framework approach to link levels of performance and report on a study which used FACETS to establish vertical linking of the three CELS writing tests.

This issue carries two articles reporting on studies relating to speaking test validation; both contributions come from research students currently undertaking PhDs in language testing and with whom we are collaborating in a variety of ways. Yang Lu, at Reading University in the UK, is currently exploring the nature and assessment of discourse competence in the spoken language using data from a subset of FCE Speaking Tests; her analysis provides some empirical support for the discourse competence assessment criterion widely used in the Cambridge ESOL Speaking Tests. Lindsay Brooks is at OISE at the University of Toronto, Canada, and has been working with us to develop and apply an Observation Checklist within the ongoing validation programme for the recently revised IELTS Speaking Test.

The last quarter of 2002 proved a busy conference season for various staff within the Research and Validation Group. In this issue we have included reports on the BAAL, LTF and LTRC conferences, as well as news of upcoming conferences which are of particular interest to applied linguists and language testers later this year.

Finally, we include a section relating to the IELTS MA Dissertation Award, with a photograph of the presentation to the 2001 award winner, which took place at LTRC 2002 in Hong Kong, as well as details of submission procedures for the 2003 award.

Published February 2003

Responding to diversity: providing tests for language learners with disabilities

LYNDA TAYLOR, CAMBRIDGE ESOL RESEARCH AND VALIDATION GROUP
MIKE GUTTERIDGE, CAMBRIDGE ESOL SPECIAL NEEDS CONSULTANT

Introduction

There has been considerable debate in recent years about the ethical dimensions of testing and assessment; this is an area we have touched upon in previous issues of *Research Notes* in relation to our role as an international examination provider. One important measure of the ethical standing of any examination board must surely be the extent to which it acknowledges and provides for the particular needs of candidates in 'minority' groups or with special requirements. One such group – though still highly diverse in its nature – is the population of candidates with disabilities, whether permanent or temporary.

Educational, employment and social opportunities for people with disabilities have increased steadily over recent years; this growth in opportunities, combined with a greater awareness of individual human rights, has understandably led to increased demand for testing and assessment provision. Language test providers need to be able to offer special arrangements for test-takers with disabilities (those with learning disabilities as well as those with visual, hearing or physical impairments); providing such special arrangements usually involves departing from the established testing protocol and modifying test content or administration processes in some way so as to minimise the impact of test-taker attributes which are not relevant to the ability that is being measured.

This article outlines the range of special arrangement provision – sometimes known as test 'accommodations' – currently offered by Cambridge ESOL through the work of its Special Circumstances Unit and gives details of the take-up of this provision in recent years. We shall also consider some of the theoretical and practical challenges for us as test developers and highlight the role which the Research and Validation Group is playing in this area.

The Cambridge ESOL Special Circumstances Unit (SCU)

The Cambridge ESOL Special Circumstances Unit is responsible for dealing with applications for special arrangements and special consideration in respect of Cambridge ESOL products; it also deals with cases of malpractice. These three areas may be defined more precisely as follows:

- **Special Arrangements:**
  Special Arrangements are made for candidates with special needs before an examination is taken so that, as far as possible, they are then able to take the examination on an equal footing with other candidates. For example, candidates with a permanent disability, such as hearing/sight impairment, dyslexia or a speech impediment; or short-term difficulties (for example, a broken arm) may need arrangements such as modified papers, readers or amanuenses, or extra time.
- **Special Consideration:**
  Special Consideration is given to candidates who are affected before or during an examination by adverse circumstances. Examples include illness, bereavement or circumstances affecting the conditions under which an exam is taken. Special Consideration is applied for **after** the candidate sits an examination.
- **Malpractice:**
  Malpractice (defined as any conduct which has the intention, or effect, of giving an unfair advantage to one or more candidates) is brought to the attention of Cambridge ESOL via reports from Centres, reports from Examiners, and through routine statistical checks applied to candidates' answer sheets.

The following table provides a snapshot of the work of the Unit over the past three years and lists the total number of candidates applying for special arrangements, special consideration and those involved in cases of malpractice for all Cambridge ESOL examinations. The figures relate mainly to FCE, CAE and CPE (the Upper Main Suite of Cambridge ESOL examinations) as the majority of cases dealt with by the Unit involve these examinations.

Table 1: Candidates applying for special arrangements or consideration and malpractice cases 1999–2001.

	Special Arrangements (candidates)	Special Consideration (candidates)	Malpractice Cases (each case may involve one or more candidates)
1999	656	5941	90
2000	948	6441	120
2001	1135	11646	122

Although these numbers may appear small in comparison to the total test-taking population (now well over 1 million candidates annually), it should be remembered that all the above cases are dealt with on an individual basis. For example, a blind candidate may require separate facilities for taking the examination, a specially modified question paper, an individual invigilator and/or reader/amanuensis on the day, and extra time to complete their papers.

The same need for individual (and often lengthy) attention applies to applications for special consideration and, of course, in all cases where candidates have been reported for malpractice.

Candidates applying for Special Arrangements

The number of applications for special arrangements for all candidates taking Cambridge ESOL Upper Main Suite examinations (FCE, CAE and CPE) in all categories (including extra time) were as follows:

Table 2: Special Arrangements for UMS papers 1999–2001.

1999	2000	2001
561	670	966

The following table gives a brief overview of important categories of special arrangements for FCE, CAE and CPE agreed in 2001, together with comments on the total number of candidates who had these special arrangements.

Table 3: Main categories of Special Arrangements for UMS papers 2001

	March 2001	June 2001	Dec 2001	Total 2001
Braille Versions of papers	2	18	7	27
Enlarged Print papers	0	29	15	44
Lip-reading Versions of listening papers	1	27	19	47
Special Needs Versions of listening papers	3	31	35	69
Separate Marking of writing papers (dyslexic candidates)	0	64	114	178
Exemption from listening or speaking components	0	5	2	7
Total	**6**	**174**	**192**	**372**

Braille Versions

There was an increase in blind candidates applying for Braille versions of FCE, CAE and CPE in 2001 (20 candidates in 2000). The majority of applications were for uncontracted Braille – a version of Braille in which there is a separate Braille symbol for every letter, compared with contracted, where a single symbol may represent a group of letters.[1]

Enlarged Print Versions

For most examinations, these are available in either A3 or A4 format. A3 papers are the standard papers enlarged to A3 size. In A4 format papers all text is enlarged, usually to 16 point, and printed in bold. There are also changes to layout. In 2001 the majority of applications were for A4 format question papers, which are probably more appropriate for most partially-sighted candidates because of the standardisation of font size, layout, etc.

Hearing-impaired (lip-reading) Versions

There was an increase in numbers applying for special lip-reading versions of FCE, CAE and CPE Listening Tests in 2001 (up from 41 candidates in 2000).

Special Needs Listening Test Versions

A smaller number of candidates applied for these specially recorded versions than in 2000 (75 candidates). They are available to blind, partially-sighted, physically disabled, and temporarily disabled candidates (i.e. in any circumstance where a candidate is unable to write notes/answers while they are listening).

Separate marking for candidates with Specific Learning Difficulties

Candidates with Specific Learning Difficulties (such as dyslexia) can apply to have their written work marked 'separately' (i.e. with spelling errors being disregarded). An increased total of 178 candidates applied for these provisions in FCE, CAE and CPE, compared with 82 candidates in 2000.

Exemption from Listening or Speaking components

This arrangement is particularly useful in cases where candidates have severe speaking or listening difficulties. Candidates applying for this arrangement receive a certificate endorsement or 'indication' if an overall passing grade is achieved. Relatively small numbers of candidates applied for exemption from FCE, CAE or CPE Listening and Speaking Papers in 2001, and there were fewer candidates than in 2000 (11 candidates).

Performance of Special Needs Candidates

It is interesting to analyse candidate results over the three UMS sessions (March 2001, June 2001, and December 2001) for various categories of special needs candidates. Table 4 shows the number of blind candidates taking and passing Braille versions of UMS papers during 2001.

Table 4: Candidates taking UMS Braille Versions

Braille Versions: FCE, CAE and CPE	Total Candidates	Passing: (Grade C and above)
March 01:	2	2
June 01:	15	8
December 01:	7	2

In 2001, 12 out of 24 blind candidates who took UMS syllabuses achieved an overall grade of C or higher.

A slightly smaller proportion of candidates who took a lip-reading version of the Listening Paper obtained an overall pass grade.

1. Grade 1 – or uncontracted Braille – consists of 63 symbols made up of all the possible variations of a series of six dots. Twenty-six of these represent the letters of the alphabet and others represent punctuation marks. These symbols can be used to reproduce a letter-by-letter copy of print. Grade 2 – or contracted Braille – was developed to reduce the size of books and to make reading quicker. It uses combinations of symbols to represent common letter combinations or words. Some characters may change their meaning depending on how they are spaced.

Table 5: Candidates taking UMS Lip-reading Versions

Lip-reading Versions: FCE, CAE and CPE	Total Candidates	Passing: (Grade C and above)
March 01:	1	1
June 01:	27	12
December 01:	19	10

Candidates applying for Special Consideration

Table 6 shows the total number of candidates applying for special consideration in 1999–2001 for FCE, CAE and CPE. Applications processed by the Cambridge ESOL Special Circumstances Unit doubled between 2000 and 2001.

Table 6: Candidates applying for Special Consideration 1999–2001

	1999	2000	2001
March	Not Available	153	166
June	3186	3255	4860
December	2434	2211	6020
Total	**5620**	**5619**	**11,046**

Applications for special consideration are received for a wide variety of reasons. Candidates may have been affected before an examination by personal illness, accident or bereavement; alternatively, they may have been affected by adverse circumstances during the actual taking of the examination, e.g. unexpected noise, equipment failure, or some other disruption. Problems in Listening Tests, particularly with equipment, always account for a large number of applications. Other common problems include external or internal noise/disruption, problems with acoustics and, more recently, mobile phones ringing.

Appropriate action to compensate candidates affected can only be taken if the nature of the problem is accurately described by examination supervisors and staff at the test centre. Action taken depends on the type and severity of the circumstances reported.

The figures given above highlight the increasing numbers of applications for special arrangements and special consideration relating to Upper Main Suite examinations. Smaller numbers of applications for candidates taking Lower Main Suite (KET, PET) and other Cambridge ESOL products were also received by the Cambridge ESOL Special Circumstances Unit during 2001. These numbers are also growing.

Theoretical considerations and empirical investigation

Clearly, the implementation of special arrangements and special considerations raises important theoretical, practical and ethical considerations, in particular:

- how to determine a disability requiring test modification or circumstances requiring special consideration;
- what type of special arrangement to provide or what sort of special consideration to give;
- how to interpret test scores produced under special conditions.

Professional judgement clearly plays a key role in decisions about the nature and extent of modified tests. But the role of professional judgement is complicated by the fact that empirical studies in this area are often lacking due to the practical constraints of research in this field, e.g. small sample size, non-random selection of test-takers with disabilities.

It is for this reason that Cambridge ESOL is engaging in an ongoing programme of activity related to special arrangements and special considerations. Issues of current interest include:

- the assessment of writing performance by second language learners with dyslexia;
- the role of assistive technology in testing second language learners;
- policy on the use of British Sign Language in ESOL examinations;
- the training needs of oral examiners for speaking tests and of other staff involved in modified tests;
- the question of what constitutes acceptable medical evidence in support of requests for special arrangements (in an international context).

This programme of activity includes small-scale empirical studies initiated and managed by the Research and Validation Group. One aspect of this is the development of a small corpus of performances from special circumstances candidates in our examinations which can provide the necessary data for investigation of these issues.

Conclusion

The area of Special Circumstances is complex precisely because a balance is required between allowing candidates with disabilities arrangements enabling them to be placed on an equal footing with other candidates but not advantaging them to the extent that the assessment objectives of the examinations are compromised. Findings from the proposed studies should help test providers design modified tests which are better suited to the needs of test-takers with disabilities and give them the opportunity to participate in mainstream academic, professional and public life.

2003 is the European Year of People with Disabilities (EYPD) the aim of which is to raise awareness of the rights of disabled people to full equality and participation in all areas. Cambridge ESOL is keen to contribute and we look forward to reporting further on our work in this field in future issues of *Research Notes*.

For further information about Special Arrangements please see the support pages on the Cambridge ESOL website:
http://www.CambridgeESOL.org/support/specials/dyslexia.cfm

Further details of the EYPD are available at:
http://212.113.82.54/eypd/index.jsp

Producing Modified Versions of Cambridge ESOL Examinations

RUTH SHUTER, CAMBRIDGE ESOL SPECIAL CIRCUMSTANCES CO-ORDINATOR

Introduction

Candidates with visual or hearing difficulties may need adapted versions of examination material to reduce the effect of the disability on their opportunity to show their ability in English in Cambridge ESOL examinations.

It is important that these adapted versions cover the same assessment objectives as their 'standard' counterparts, and so the adapted versions are as far as possible based on the standard papers once they have been passed for print, with minor changes to rubrics, layout and sometimes length. Occasionally it is necessary to completely replace a question or item (for example, a writing task asking candidates to describe a favourite picture would be impossible to answer for a person who had been blind from birth).

Enlarged print papers are produced for visually impaired candidates for most examinations, and we also produce print versions of Braille papers for blind candidates, which are then sent to an external agency for Brailling. Lip-reading versions of listening papers are available for hearing impaired candidates for most examinations.

Which papers are adapted and when?

Adapted versions of the Upper Main Suite (UMS) papers for all syllabuses are produced routinely for all sessions because experience indicates that many of these are likely to be required, and because most UMS papers are released after each session. Off the shelf versions are produced for IELTS, Lower Main Suite, BEC and BULATS examinations. When sufficient notice is given and the material in the paper in question is suitable, adapted versions of CELS are produced on request. When sufficient notice is given, lip-reading versions of YLE listening tests are also produced on request.

In 2001 Cambridge ESOL produced a total of 137 special needs 'papers' comprising 276 items including listening tapes, contracted and uncontracted versions of Braille papers (see footnote on page 3), separate booklets for the texts and question booklets of reading papers, supervisor's booklets for the listening tests and examiner's instructions for the speaking tests.

What is the general procedure for adaptation?

A flow chart illustrating the main stages in the production of the print versions of UMS Braille and A4 enlarged print Reading, Writing and Use of English papers (papers 1–3) is shown alongside. Special needs, A4 enlarged print and lip-reading versions of the listening tests are also produced.

Figure 1: Producing UMS Papers 1–3 in Braille/enlarged print

Time	Action carried out by	
6–7 months before paper despatch	Special Needs Consultant (SpNC)	Overview of the standard papers to check for possible problems in content, etc
7 months before paper despatch	Subject Officers	Standard papers passed for print
5–7 months before paper despatch	QPP (question paper production unit) → SpNC	Papers sent to SpNC for modification
2 weeks	SpNC → QPP → typesetting	Modified papers sent to typesetting
2 weeks	Typesetting → QPP → SpNC	Papers set, first proofs sent to SpNC for checking
10 days	SpNC → QPP → Special Circumstances Co-ordinator (SpCC) → QPP → typesetting	First proofs sent to SpCC for checking and revision if necessary
4 weeks	SpCC → proof reader → SpCC	SpCC sends revised proofs to external proof reader for final check
1 week	SpCC → typesetting	SpCC checks comments, makes any final amendments
4 weeks	SPCC → QPP	Papers passed for print. Time between approval for print and despatch includes time for print versions of Braille papers to be Brailled. Also allows extra time for more revisions to A4 enlarged print papers if required (format changes to A4 enlarged papers tend to mean more revisions are required than for print versions of Braille papers). Time for printing.
3 weeks		
Ideally 4 weeks before exam		Papers despatched to centres.

What about listening tests?

The production of the listening tests starts a month or so before that of other papers. The Special Needs Consultant checks the different test versions available for each qualification, and recommends which would be the best version for adaptation. Criteria which are taken into account when choosing which version to adapt for the Special Needs/enlarged print versions include:

- Content which would be inaccessible to blind candidates;
- Task types which might be confusing to blind candidates (e.g. all other things being equal, a multiple matching task is more difficult for a blind candidate than a multiple choice task, because blind candidates find it more difficult to get an 'overview' of the whole task);
- A version in which the amount of text in the questions is lower than in other versions is preferable;
- If all the criteria above are equal, it is preferable to adapt the version of the standard listening test which is released for general use, as this increases the amount of special needs practice material available.

For lip-reading versions for UMS and other higher level qualifications, monologues or prompted monologues which can be 'turned into monologues' are needed, as it would be too difficult for a seriously hearing impaired candidate to try to follow two supervisors reading different parts. This may mean that it is necessary for some tests to include material from more than one standard version. If this is the case, every effort is made to ensure that the material used as a replacement has similar characteristics to the original material it is replacing.

The special needs and enlarged versions of the listening tests involve recording a new rubric and producing a supervisor's booklet. The booklet is based on the studio rubric used for the recording of the test, with the additional inclusion of the tapescripts for the different Parts. These are marked with asterisks, which indicate where the supervisor should pause the tape during the second hearing of each Part to allow the candidate time to read the questions and write or check the answers. The special needs version tapes are recorded approximately 6 months in advance of the first possible despatch date.

The special needs versions of the listening tests are used for candidates who are unable to write at the same time as they listen, due to, for example, a broken arm. They are also the versions sent to be Brailled. Enlarged print papers are also produced for the same tapes and supervisor's booklets.

The lip-reading versions of the tests comprise a supervisor's booklet, which again includes the tapescripts, marked with asterisks to show where the supervisor should pause during the second of the three readings of each Part, and a question booklet.

And speaking tests?

Three or four of the packs available for each qualification are adapted for use by visually impaired or blind candidates each year. The choice of which packs are most suitable for adaptation is made by the relevant Subject Officer, the Special Needs Consultant and the Special Circumstances Co-ordinator 6–9 months ahead of the first possible despatch date.

Blind candidates are given a Braille description of the visual stimuli for each task. Visually impaired candidates can choose to use either an enlarged print version of the written descriptions used by blind candidates, or can ask to be supplied with enlarged versions of the visuals themselves.

Criteria which are taken into account when the choice of packs is made include:

- Overtly visual tasks such as 'Which picture would be the best for a telephone card?' are unsuitable;
- It must be possible to adequately capture the essence of the picture in a fairly short description which makes it possible for the candidate to carry out the task;
- The description should not supply the candidate with structures or vocabulary at a level which would be considered to be at or above the level for the qualification in question;
- It must be possible to describe the picture without partially completing the task for the candidate (for example, tasks which ask candidates to speculate on how people are feeling are rarely suitable as the description would have to answer the question);
- These materials are also used for candidates taking the test in a single format in prison, so it is not a good idea for all of the choices to involve pictures of children or family life, for example.

Oral Examiners conducting Speaking Tests for Main Suite candidates are supplied with the following:

- General advice on conducting tests for candidates with various types of difficulty;
- Specific instructions on how to conduct the tests;
- Adapted interlocutor frames for candidates taking a single format test;
- Print copies of Braille descriptions.

Conclusion

Producing modified materials for candidates taking Cambridge ESOL examinations is a long, involved process, and in order to ensure that the maximum number of candidates benefit from this provision it is important that as much notice as possible is given when candidates need modified papers. We are also looking at ways of streamlining the process and making it more efficient, while making sure that modified papers are produced to the same rigorous standards as the standard papers.

Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts

STUART D SHAW, CAMBRIDGE ESOL RESEARCH AND VALIDATION GROUP

It is widely held that handwriting, neatness and layout contribute to the basis of legibility. A deeply entrenched conventional wisdom coupled with intuitive belief suggest that handwriting affects the assessment of a piece of extended writing. Accordingly, Hughes, Keeling and Tuck (1983) have argued that raters with neat and presentable handwriting significantly underrate untidy and illegible written responses. Other things being equal, well-presented constructed responses tend to receive higher scores than their poorly-presented counterparts (Chase 1968; Briggs 1970; Markham 1976).

The introduction of computer administered direct tests of writing – in which examinees can choose to word-process their responses – has raised fundamental considerations regarding salience of legibility and the rating of second language writing. Clearly, in translating a test from one medium to another, new medium it is crucial to ascertain to what extent the new medium may alter the nature of the underlying test construct, or change the scale. Some of the specific research considerations in relation to this include:

- the impact of composition medium on essay raters in second language writing assessment;
- the significance and impact of the role of legibility in the assessment of word-processed scripts;
- whether raters rate handwritten and word-processed responses differently and, if they do, whether any differences interact with gender, ethnicity or socio-economic background.

Brown asserts that "handwriting and neatness has long been seen as a contaminating factor in the assessment of writing ability" (2000:1). Throughout recent years there have been several studies in the area of first language writing assessment which have attempted to investigate the impact of legibility on the general judgement of writing quality. In the main, the quality of handwriting has an effect on the scoring of essays with improved legibility resulting in higher awards. In contrast, however, there exists a paucity of studies examining the effect of handwriting in the assessment of second language writing.

One of the few *second* language investigations, conducted by Robinson (1985), was able to replicate first language assessment findings; namely, that essays written by students whose first language did not employ the Roman alphabet tend to be awarded lower scores than essays composed by 'expert' writers.

Assessment constraints in the language testing context – multiple assessment foci and restricted time – have, according to Charney (1984), resulted in handwriting playing a more significant role in assessment than it perhaps should. With the general requirement to read and rate many essays rapidly, raters are being compelled to "depend on those characteristics [such as handwriting] in the essays which are easy to pick out but which are irrelevant to 'true writing ability'" (Charney 1984).

Chou et al (1982) have suggested that essays are easier to read if presented neatly and that it is not only hand writing *per se* that creates a favourable impression in the mind of the rater but that severe text editing may produce an unfavourable impression. The implication is that not only is poor handwriting difficult to process but that raters, on the basis of script untidiness, forge rather negative inferences about the character and personality of the author. Dramatic examples of script revision may be negatively interpreted by raters as being indicative of a candidate wholly ill-prepared for writing, devoid of any sense of effective textual organisation.

Huot (1993) puts forward the argument that fast reading during examination marking would be expected to be affected by handwriting: untidy and illegible handwriting is likely to impede fluent, rapid reading especially in second language contexts where the centrality of 'fluency' is a major component of communicative quality. Vaughan (1991), investigating protocol analysis as a means of identifying factors which influence the assessment of second language writing, concluded that handwriting and overall presentation are especially important to raters on the basis that the number of direct references to handwriting was second only to aspects of content. Milanovic et al (1996:106) seem to support this finding by suggesting that layout appears to engender particular prejudices in certain raters before they even consider the content of a response. This also applies, it would seem, to the recognition of handwriting across national group (e.g. French or Italian which have common features across writers) which lead to observations which indicate some effect on rater judgement.

A recent study by Brown (2000) which investigated differences between handwritten and word-processed versions of the same IELTS Task Two essays and the effects of handwriting on legibility and assessment, deduced that legibility has a marginal but significant impact on scores. Moreover, the size of the effect is relative to the quality of handwriting and neatness of presentation. Contrary to her hypotheses, the handwritten versions of the same script were assessed higher than the word-processed versions. Therefore, in this study, the worse the handwriting, the higher the comparative assessment. A secondary aim of the study was concerned with the validity of verbal protocols as evidence of raters' assessment orientation when marking essays. Protocol

analyses of the raters revealed that they may well have been compensating for poor handwriting in their assessment. Whilst the proportion of pejorative comments increased with regard to the reduced legibility of responses, the ratings also increased, rather than decreasing as might be predicted. The study underlined the problems associated with assuming the veracity of raters' observations as accurate indicators of rating orientation.

Investigation of FCE hand-written and word-processed scripts

An investigation of FCE (0100 syllabus) June 2000 hand-written and typed versions of the same scripts was undertaken. The study aimed to deduce whether salience of legibility, as realised through quality of handwriting, contributes to an understanding of rater bias and focused on two key questions:

- What is the impact of legibility on ratings awarded to FCE handwritten and word-processed Writing tasks?
- How are raters affected by aspects of legibility and presentation as manifested by handwritten and word processed responses?

Three highly experienced, current FCE examiners participated in the study. Each examiner independently re-rated 75 scripts typed up from their original handwritten forms. Examiners were additionally asked to comment on the marking experience by completing a questionnaire.

Script preparation

Details of candidate name and centre were removed from the scripts so as not to unduly influence the raters and the remaining text was word-processed. Keyers typed exactly what appeared on the original responses with no changes whatsoever. The examiners were, therefore, provided with an exact representation of each script. Letter case was keyed as in the original script as was the punctuation used and the format of the script as far as possible. Each script was presented separately – by task – on several pieces of paper. These versions were given to each of the three examiners. It should be noted that the typed-up scripts were not 'authentic', in the sense that they were not produced on the keyboard by the candidates themselves under examination conditions although they were accurate reproductions of writing done under such conditions.

Assessing FCE scripts

Raters were provided with two sets of band descriptors: an impression mark is awarded to each piece of writing using a general mark scheme which is used in conjunction with a task-specific mark scheme, which focuses on criteria specific to each task. To rate a script, FCE examiners follow a prescribed scoring method and decide which band descriptor most accurately describes the candidate's performance. An appropriate band score is subsequently awarded. Each script contains two tasks, the compulsory Task 1 and Task 2 which offers candidates a choice between several options. The scripts selected for this study all contained the same Task 2 response.

Study findings

The impact of rating typed versions of the original handwritten scripts is to deflate the mean – a finding which is in line with current research (Brown 2000). In both Task 1 and Task 2, the mean is lower for the typed texts than for their handwritten counterparts. At first glance, the direction of this effect might be unexpected i.e. increased script legibility might be thought to produce higher scores and poor legibility thought to lead to lower scores – as is the case for first language assessment. However, in second language writing assessment, it may be the case that greater stress is put on certain linguistic features such as grammatical accuracy and range, lexical resource and syntactical structures. Unlike first language assessment, mastery of orthographic and iconic conventions and handwriting neatness may not be particularly significant assessment foci (Cumming 1998). In second language assessment there is instead a marked central focus on mechanical aspects of writing. Poor legibility might well serve to distract from mechanical errors. Second language raters may even attempt to compensate for poor legibility by more careful and attentive reading in the hope of avoiding discriminating against candidates because of poor script legibility.

Table 1 shows the mean scores by task for each rater and the original score.

For both Task 1 and Task 2, standard deviations for handwritten texts are greater than for typed scripts implying that more of the mark scheme scale is used when applied to handwritten responses. The mark scheme applied to the compulsory task (Task 1) is both more rigid and more clearly defined than its optional task counterpart (Task 2). Additionally, the range of language required by the compulsory task is less wide and its focus is discursive whereas the optional task permits more scope for invention and a variety of interpretation. Consequently, examiners are allowed greater freedom in their assessment of the optional response as demonstrated by the higher standard deviation indices for Task 2 for both handwritten and typed scripts.

In order to investigate the significance of the score differences for the two script types for each rating category, a Matched t-test and a Wilcoxon matched-pairs signed-ranks test was carried out. The re-ratings were averaged and compared with the original ratings in order to create pairs for significance testing. The z-score statistics for the Wilcoxon test confirmed the matched t-test findings that ratings differ significantly between handwritten and typed scripts. We can have some confidence in concluding that typed scripts do have an effect on the way that scripts are rated.

Correlational analyses were used to investigate the relatedness of the ratings for the handwritten and typed texts. A useful way of interpreting a correlation coefficient (r) is to convert it into overlap between the two variables. This permits an appreciation of how much of the variance (the variability of scores around the mean) in

Table 1: Mean Band Scores by Task

Rater	Task 1			Task 2		
	Mean	Std.Dev.	Range	Mean	Std.Dev.	Range
RATER 1	2.8853	0.91144	3.90	2.9747	1.14669	4.10
RATER 2	3.2680	0.89309	3.10	3.4653	1.09341	4.00
RATER 3	3.0867	0.94959	4.00	3.3013	1.18886	4.10
RATER Av	3.0800	0.91804	3.67	3.2471	1.14299	4.07
ORIGINAL	3.3827	1.03615	3.20	3.3347	1.24635	4.10

one measure can be accounted for by the other. To the degree that the two measures correlate, they share variance. The overlap i.e. the square of the correlation coefficient, informs us that the measures are providing similar information. In other words, the magnitude of r^2 indicates the amount of variance in one set of marks (original ratings) which is accounted for by the second set of marks (re-ratings) or vice versa. In order to show that handwritten and typed ratings measure essentially the same thing, a correlation coefficient in the high 0.80s or 0.90s would be required, i.e. a very high overlap of the two measures (Hatch and Lazaraton 1991:441). In this study, the correlational indices are somewhat lower, shown in Table 2.

Table 2: Correlational Indices

Re-ratings	Original ratings	
	Task 1	Task 2
Rater 1	0.754	0.816
Rater 2	0.830	0.833
Rater 3	0.726	0.769

Table 3: Inter-rater reliabilities

Inter-rater reliabilities		
Task 1	Task 2	No. Scripts
0.745	0.785	75

In order to confirm that the rating data is reliable, it is necessary to calculate the inter-rater reliability for the three trial raters. To compute the inter-rater reliability a Pearson correlation matrix was generated and then an average of all the correlation coefficients was derived. Any distortion inherent in using the Pearson for ordinal data was corrected for by applying a Fisher Z transformation to each correlation. Inter-rater reliabilities between trial examiners on both tasks were encouragingly high – this level of reliability can be considered acceptable.

Effect of rating word-processed and handwritten scripts

All three examiners were enthusiastic about the marking trial claiming it to be a positive experience. FCE examiners, it was believed, would appreciate not having to read poor handwriting. Despite initial difficulties, once examiners had gained familiarity with the typed responses they became increasingly more comfortable assessing them. Furthermore, the speed of assessing scripts increased with time and examiners achieved their normal marking rates.

Examiners identified a number of advantages when marking type-written scripts:

- both strong and weak scripts are easier to read. One examiner remarked:
  "*We often wonder if a candidate would have been given a better mark if we had been able to read it or read it easily for there are few that are really illegible but many that look a mess and are hard to decipher. These tend to get lower marks for language control*".
- poor handwriting is not penalised. Raters appear not to be unduly influenced by neatness of presentation which is exhibited through handwriting;
- errors of spelling and punctuation are accentuated when typed and are more easily identifiable;
- all scripts look similar in general appearance before reading thereby facilitating rater objectivity;
- typed texts facilitate paragraph identification.

However, examiners did point out certain disadvantages associated with rating typed scripts:

- it is often hard to assess the length of a response especially in regard to word count (as type-written scripts have less space between words);

- it is easier to miss spelling errors or to wrongly penalise scripts for poor spelling.

Asked how they would assess their own marking of typed scripts in relation to marking hand-written scripts, one examiner believed that she was more objective and thorough when assessing typed versions:

" … not knowing on first impression that a candidate was either very young or too familiar with the script".

Another was more concerned with script throughput and although her first 50 scripts took longer to assess than normal, she soon achieved normal marking rates. The third examiner was primarily concerned with accuracy and felt that it was harder to ascertain how accurate she had been as she was unable:

"to form a picture after marking up the script than with handwritten ones (but then I'm not used to these)".

Examiners agreed that it was difficult not to be influenced by bad handwriting and any decisions which needed to be taken in order to relegate an assessment from one band to a lower band were very subjective. Interestingly, whilst one examiner believed that when she began the trial marking process she was heavily influenced by neat type rather than by linguistic content, she later concluded that there appeared to be no difference between the two forms of script.

Conclusions of the study

The effect of improved legibility, neatness of presentation and layout – as realised through typed responses – does not seem to favour typed scripts. Indeed, the impact of rating typed versions of the original handwritten scripts is to deflate the mean. The strong emphasis, in the assessment of second language writing, on linguistic features might account for this finding.

Whilst examiners feel that it is difficult not to be influenced by bad handwriting they also consider that consistent format aids assessment of language and task and that a typed response can engender an objective approach to rating. Examiners were unanimous in their belief that typed scripts were easier to read than handwritten scripts, paragraphs were more readily located and spelling/punctuation errors were accentuated permitting their immediate recognition.

This study has revealed insights into how examiners approach and rate different forms of candidate writing. It has implications both for the rating of Special Arrangements candidates' writing (i.e. those with Separate Marking or those who type their answers) and for aspects of computer-based testing and electronic marking of scripts, two technological areas which will be reported in the next issue of *Research Notes*.

References and further reading

Bridgeman, B and Cooper, P (1998): *Comparability of Scores on Word-Processed and Handwritten Essays on the Graduate Management Admissions Test*, paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, April 13–17 1998.

Briggs, D (1970): The influence of handwriting on assessment, *Educational Research*, 13, 50–5.

Brown, A (2000): *Legibility and the rating of second language writing: an investigation of the rating of handwritten and word-processed IELTS Task Two essays*, IELTS Research Projects 1999/2000.

Charney, D (1984): The validity of using holistic scoring to evaluate writing: a critical overview, *Research in the Teaching of English*, 18, 65–81.

Chase, C I (1986): Essay test scoring: Interaction of relevant variables, *Journal of Educational Measurement*, 23, 33–42.

Chou, F, Kirkland, J S and Smith, L R (1982): *Variables in College Composition*, Eric Document Reproduction Service No. 224017.

Hatch, E and Lazaraton, A (1991): *The Research Manual: Design and Statistics for Applied Linguistics*, Boston, Mass: Heinle and Heinle Publishers.

Hughes, D C, Keeling, B and Tuck, B F (1983): Are untidy essays marked down by graders with neat handwriting? *New Zealand Journal of Educational Studies*, 18, 184–6.

Huot, B (1993): The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends, *Review of Educational Research*, 60 (2), 237–263.

Manalo, J R and Wolfe, E W (2000): *The Impact of Composition Medium on Essay Raters in Foreign Language Testing*, paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 24–28 2000.

Markham, L R (1976): Influences of handwriting quality on teacher evaluation of written work, *American Educational Research Journal*, 13, 277–83.

Milanovic, M and Saville, N (1994): *An Investigation of Marking Strategies using Verbal Protocols*, UCLES 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.), *Studies in Language Testing – Performance Testing, Cognition and Assessment; Selected Papers from the 15th Language Testing Research Colloquium at Arnhem*, Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press, pp.92–114.

Robinson, T H (1985): *Evaluating foreign students' compositions: the effects of rater background and of handwriting, spelling and grammar*. Unpublished PhD thesis, The University of Texas at Austin.

Vaughan, C (1991): Holistic Assessment: What goes on in the rater's mind? in Hamp-Lyons, L (Ed.), *Assessing Second Language Writing in Academic Contexts*, Norwood, NJ: Ablex Publishing Corporation, pp.111–125.

Task difficulty in the assessment of writing: Comparing performance across three levels of CELS

NEIL JONES AND STUART D SHAW, CAMBRIDGE ESOL RESEARCH AND VALIDATION GROUP

Introduction

CELS (Certificates in English Language Skills) is a suite of assessments introduced in May 2002. It is a modular system which enables candidates to enter for one or more skills (Reading, Writing, Listening and Speaking) and to choose the level at which to enter (Preliminary, Vantage or Higher, corresponding to ALTE levels 2, 3, 4, or Council of Europe Framework (CEF) levels B1, B2 and C1). Like the examinations it replaces – Certificates in Communicative Language Skills (CCSE) and the Oxford EFL examinations – it emphasises communicative language ability and the use of authentic materials. While attempting to maintain continuity with the CCSE and Oxford EFL examinations, the development of CELS links it firmly to the Cambridge ESOL framework of levels, and thus to the corresponding Main Suite exams at the same levels (PET, FCE, CAE). This linking is achieved for the objective papers (Listening and Reading) through the use of Main Suite anchor items of known difficulty in the pretesting of CELS, and through other specific studies. For the performance papers (Speaking and Writing) the linking comes from a common approach to markscheme design, training and standardisation, verified again by some specific re-marking studies (see *Research Notes 9*).

Exams which test at a level, with passing and failing grades, are particularly attractive in the context of language study, within a school or some other setting. The exam can provide material and elicit performance which is appropriate to the level, and which can impact positively on the learning process. However, this can also raise issues for the vertical comparison of lower with higher levels. In the case of Writing it has been recognised that the setting of different kinds of task, with different degrees of challenge, complicates the comparison of performance across levels. This article describes a small study devised to establish empirically the relation of the three separate CELS rating scales to each other.

Background to the CELS Writing test

The CELS Handbook states that the aim of the exam is to *"assess English language competence through a variety of authentic tasks based on authentic texts"* and to test writing *"through a variety of tasks which, where possible, reflect real-life situations and test a wide range of writing skills and styles"* (2002:7). The Handbook adds, true again to the communicative writing construct, that for *"each task a context and a target reader is specified in order to establish a purpose for the writing"* (2002:37).

Morrow (1979), an early pioneer in the theory and practice of communicative testing, and other like-minded testers have suggested that performance tests should use tasks that are interaction-based, rooted in the context of a situation (with its physical environment, participants with status and roles, attitudes, and formality levels), and characterised by purposes, authenticity, and behavioural outcomes on which the performance of participants may be evaluated. The CELS Writing exams have been designed to reflect such views.

Generalisability and Performance Tests

The assessment of writing depends on the subjective rating of a specific sample of performance elicited by one or more test tasks. Tasks are selected which can be seen as appropriate to the level. The elicited sample is rated by an examiner trained to judge the performance and assess the candidate's level. A writing task needs to give all candidates opportunity to perform to their utmost abilities whilst simultaneously eliminating variations in rating that can be ascribed to the task rather than the candidates' respective abilities. This raises two fundamental issues:

1. On what dimensions do writing tasks vary, both in language testing situations and in the 'real world'? Such issues relate to content coverage or 'construct representation' (Messick 1989): language testers are especially interested in sampling from a specific domain of writing in a writing test and it is, therefore, useful to first of all describe the domain;
2. Given the many ways in which writing tasks can vary, which of these ways are associated with different levels of candidate performance and, equally important, which are not? This question is relevant for several reasons:
   - It is necessary to minimise the degree of both random and systematic error in the test – referred to as 'construct-irrelevant variance' (Messick 1989). Comparability is easier if test takers interpret the task in the same way.
   - As a test of writing may consist of several tasks, and may even allow a choice between tasks, it is important to ascertain to what extent the task or prompt can be varied and still produce comparable results;
   - If different kinds of task elicit different kinds of performance, we should know wherein these differences lie. Are they associated with observable differences in the grammatical and syntactical, lexical or rhetorical features of texts, or can differences in ratings be primarily attributed to certain aspects of the rating approach – in other words, do raters employ different criteria in assigning ratings to different task types?

We wish to be able to describe levels of writing proficiency in terms which are as generalisable as possible, while recognising that there is an inevitable conflict between the general and the specific.

Moving from description to assessment the problem is even more evident. Given that performance tests attempt to reflect specific authentic communicative tasks, they inevitably encounter problems of '*generalisability to other performances and extrapolation to future abilities'* (Hawkey forthcoming). It has been noted that specificity of task tends to increase difference in task effect on performance (Bachman et al 1995, Hamp-Lyons 1995).

Weir (1993:11) perceives rigour in the specification of direct performance tasks as one possible way to increase generalisability. The sample of communicative language ability selected for a test must be *'as representative as possible'*, the test tasks in accordance with *'the general descriptive parameters of the intended target situation particularly with regard to the skills necessary for successful participation in that situation'*. Tests should, consequently, meet 'the performance conditions' of the authentic context.

Task difficulty and performance quality

Frameworks describing levels of language proficiency are naturally expressed in terms both of what people can do and how well they can do it: that is, in terms of task difficulty and performance quality.

For example, the ALTE Can do statements for the Work area contain the following statement at Level B1: *"CAN write straightforward, routine letters of a factual nature, for example a letter of enquiry; but her/his work will require to be checked."*

The assessment of Writing would thus seem to require the rater to make a judgement based simultaneously on two things: the *difficulty* of the task and the *quality* of the performance. However, these have been characterised as quite distinct approaches to assessment, namely *counting* and *judging* (Pollitt 1991). Using a sporting metaphor, he exemplifies the counting approach with events like the high jump, weightlifting, or running. A grammar test using multiple-choice items represents the counting approach, and it is possible to estimate the precise difficulty of each such item in the test. An example of a judging event is ice skating – essentially a subjective process, like the rating of performance tests. Spolsky (1995) uses the same analogy when he asks *"whether language proficiency can be measured on a definable dimension, like the time of a race or the distance of a jump or the number of goals, or whether it must be judged on a subjective set of criteria, like the performance of a diver, gymnast, or skater"*.

In current approaches to writing assessment it turns out that task difficulty is not explicitly considered – or rather, that task difficulty is factored out of the assessment process.

To see why this is so let us first look at markschemes – the scale descriptions used for rating. It is clear that they frequently bear little resemblance to user-oriented can do descriptions like, for example, the ALTE Framework statements illustrated above. As Alderson (1991) points out, scales have different audiences (test constructors, raters, end-users) and are worded accordingly.

Although the CELS markschemes cover three levels, their wording at each level is very similar, and use is made of evaluative terms such as "inadequate", "satisfactory", "good" etc., which are to be interpreted in relation to the given level. While there are references to aspects of task fulfillment, particularly in respect of the inclusion of particular content points, the overall focus is on the *quality* of performance (in relation to the criterion level) rather than on success at completing the task.

Clearly, the real definition of the level is not captured on paper at all, but rather in the process of training and standardisation of examiners. It depends crucially on exemplar scripts, i.e. scripts which have been identified as exemplifying the level by more senior examiners. In this light, markschemes are little more than mnemonic devices for use by examiners who have already internalised a representation of the levels.

Wolf (1995) concludes that standards are communicated by examples of students' work rather than by explicit assessment criteria. However, it is not clear how exemplar scripts achieve their standardising effect:

> *"In spite of the enormous potential importance of examples and exemplars in any criterion-referenced system (indeed in any assessment system at all) there seems to be very little empirical research on their efficacy in creating common understandings and standards".* (Wolf 1995: 76)

So how does writing assessment reflect task difficulty? Features of the wording of the general markscheme may relate to task fulfilment, and may be supplemented by task-specific markschemes e.g. specifying obligatory content points. However, what the examiner is really doing is matching the sample of performance to an internalised representation of the level. This representation is the result of the training and standardisation process. The task itself is not central, in fact the approach is essentially to factor out the task and look beyond it to interpret the sample of performance in terms of the criterion level. This explains why attempts to link perceived writing task-type difficulty and writing test results have had mixed success (e.g. Hamp-Lyons and Prochnow 1991). The study reported here also confirms that raters effectively factor out task difficulty in their assessment of level.

Vertical equating of CELS

CELS writing markschemes depend upon an approach which uses examplar scripts and the training of examiners to apply an internalised standard to samples of writing performance. CELS examiners mark scripts at each of the three levels. However, there is no explicit relation of the levels in the construction of the markschemes. This is not necessarily a problem, because, as noted above, the development of CELS already locates each exam within the overarching Cambridge ESOL levels framework. However, the introduction of CELS provided an opportunity to attempt an empirical equating of three levels of writing ability.

CELS writing papers at the three levels – Preliminary, Vantage and Higher – were administered to learners of the appropriate proficiency level. Forty-five scripts – 15 at each level – were selected. Three experienced and independent examiners rated every script, using one of the three markschemes – that is, they were asked to mark each script applying the Preliminary, Vantage or Higher standard. Level 2 (Preliminary) scripts were re-marked using Level 2 and 3 rating scales, Level 3 (Vantage) scripts using all three rating scales, and Level 4 (Higher) scripts using Level 3 and 4 rating scales. Each script was thus rated using two or three rating scales, each rater used all the rating scales, and each rater rated every script in the sample.

A realistic feature of this study is that candidates responded to tasks at one level only (corresponding to their approximate proficiency level). This corresponds to the reality of testing writing – our view of a candidate's ability is always mediated by the tasks they are set. In order to achieve a link across levels, examiners were asked to use markschemes – that is, to apply their internalised representation of an exam level – to tasks set at a different exam level. This is unrealistic to the extent that it does not correspond to any operational procedure. However, it permits an empirical estimation of the actual overlap between the levels in the way that the markschemes were interpreted by these particular examiners. This overlap appears to be approximately a band per level, for example a Vantage script given a 3 using the Vantage markscheme would quite likely get a 2 using the Higher markscheme or a 4 using the Preliminary markscheme.

The response dataset was analysed using FACETS (Linacre 1988) a software package which implements a multi-faceted Rasch (MFR) approach. The Rasch model describes the probability of success on a task as a function of the difference between the ability of a person and the difficulty of the task (e.g. high ability and low difficulty means a high probability of success). The multi-faceted Rasch model extends this by enabling task difficulty to be decomposed into a number of facets. For example, a score in a writing exam might be modelled to reflect the difficulty of the task, the severity of the rater, and so on:

score = ability of the test taker – difficulty of the task – severity of the rater – level of the markscheme

The different facets of difficulty all impact on a candidate's probability of getting a good mark. In this study it turned out to be the markscheme facet which captures the vertical progression of the CELS levels. When a rater applies a markscheme, using his/her internalised representation of that particular level, then clearly, the higher the level, the lower the mark awarded for a given script. In this way the Higher level markscheme acts like a severe rater whereas the Preliminary level markscheme acts like a lenient rater. The rating scales are shown anchored to the markscheme difficulties in the FACETS output.

Figure 1 shows the output from the FACETS analysis. The vertical scale along the left of the figure represents the measurement scale, against which all the elements can be located

Figure 1: FACETS output for CELS Vertical Marking Study

```
Vertical = (1A,2A,3A,4A,5A) Yardstick (columns, lines) = 0,5
```

Measr	+script	-rater	-mark scheme	-Scores	S.1	S.2	S.3
+ 4 +	HI1	+	+	+	+ (5)	+ (5)	+ (5) +
						–-	4
	HC1 HF1						
	HC5 HF3						
+ 3 +	HC4 VH5	+	+	+	+	–-	+ –- +
	HI2						
	HF2					4	
	HC3						3
	HC2 VE4						
+ 2 +	HF5 HI3 HI5	+	+	+	+	+	+ +
	VB4 VE3 VH2					–-	
					4		–-
	VB1 VE5						
	VH4			V4			
+ 1 +		+	h +		+	3	+ 2 +
	HF4 PG1 VB5			P1 P3			
	VE2				–-		
	VH3			P2		–-	
			v	P4			–-
* 0 *	PA1 VB2	* XX *		* H6 V3 V5 *		*	* *
	PA2 PD1 VB3			H3 H4	3		
	PD3 PD4 PD5 VE1 VH1			H2 V1		2	
	PD2 PG3 PG5						
				H1 H5	–-		
+ -1 +	PA4 PA5	+	p +	V2	+	+	+ +
	PG4						
	PG2					–-	
					2		
+ -2 +		+	+	+	+	+	+ +
							1
					–-	1	
+ -3 +		+	+	+	+	+	+ +
						–-	
	HI4						
+ -4 +		+	+	+	+ 1	+	+ +
	PA3						
+ -5 +		+	+	+	+ (0)	+ (0)	+ (0) +

Measr	+script	-rater	-mark scheme	-Scores	S.1	S.2	S.3

(candidates by their ability, raters by their severity, etc). Each candidate performance (script) is represented by a code e.g. HC4 (H = Higher, V = Vantage, P = Preliminary). Candidates are ordered from most able (at the top) to the least able (at the bottom). The other facets are ordered so that the most difficult element of each facet is towards the top, and the least difficult towards the bottom. The most likely scale score for each ability level is shown in the right-most column.

A FACETS analysis can be very informative because problems which emerge in an analysis can reveal substantive issues in assessment which are reflected in the underlying response data. Two such issues arose in this study:

1. FACETS identified three major disjoint subsets. This means that the connections between certain facets are not such as to enable them to be unambiguously estimated. In this case it is the candidates who cannot be separated from the tasks. Higher level candidates responded to Higher level tasks only, and so on for each level. This gives rise to the three disjoint subsets identified by FACETS. Thus FACETS identifies as a data issue what is in fact part of the basic conceptual issue of generalizability in performance assessment: judgements of ability are inseparable from the tasks which are used to elicit performance.

2. The FACETS analysis successfully ranked the candidates by ability, the examiners by severity, and the markschemes by level. However, it failed to rank the tasks by difficulty, at least in the expected sense that Higher level tasks should be more difficult than Vantage level tasks and so on. This reflects the previous discussion of rater behaviour: what raters do in the approach to assessment used here is effectively to factor out the difficulty of the task to make a judgement about the level of the performance elicited by the task.

Exploiting task difficulty in the construction of a Common Scale for Writing

This small-scale study shows clearly that the factors of task difficulty and performance quality are practically difficult to disentangle in performance assessment. This seems to leave us in a somewhat paradoxical situation.

On the one hand, it is clear that a vertical dimension of writing ability can be constructed – that levels of writing proficiency can be plausibly described in terms of what people can do and how well they can do them. Different aspects of "can do" can be identified and described: functional, in relation to real-world tasks; linguistic, in relation to features of accuracy, range, etc. and impact on reader, in relation to those features which are salient at different levels. Such descriptions can certainly inform the design of writing assessments and choice of appropriate tasks. There is, for example, a recognisable correspondence between the Council of Europe general illustrative descriptors for writing and the tasks included in the CELS tests used in this study.

However, such descriptive frameworks do not translate in any simple way into rating instruments. Particularly when an exam tests at a single level, as most Cambridge ESOL exams do, then it seems to make sense to focus on that level in selecting appropriate tasks – there would be little point in including very easy tasks that would be completed perfectly, or very difficult tasks that would perhaps produce an uninterpretable sample of performance. But then the question becomes: does this sample of performance meet the description of this level, or not? This is a complex question, the answer to which entails finding some best fit in relation to a range of perhaps conflicting features of the performance – register, accuracy, range, impact, task fulfilment and so on. In this situation it is the use of exemplar scripts, carefully chosen by experts, and training and standardisation centred on these, which in practice guarantees the application of consistent standards across exams and across sessions. The study confirmed that, within this paradigm, it is not possible to describe the vertical progression of CELS in terms of empirically-established task difficulties.

So is there a way in which tasks, their difficulty, and interpretations of "can do" in relation to those tasks, could be brought more explicitly into the assessment of writing? This small study shows the practical problems in demonstrating that the intended level of a performance test, in terms of some encompassing framework, corresponds to its actual level. The current approach to setting and maintaining standards can be made to work well enough in practice, but it is the difficulty of parameterising task difficulty which makes the link to a framework perhaps less transparent.

Considering possible developments for CELS there are issues concerning test design. For example, could a common task be included in tests at adjacent levels? This would have face appeal as an empirical basis for vertically linking the CELS levels, but as this article has shown, would not necessarily work.

There are also considerations for markschemes:

- Could a single markscheme covering three levels be used, or could explicit indications of equivalence across levels be given?
- Should markschemes use more specific can do statements?

The issue here is whether any changes would enable examiners to rate more consistently, and accurately in relation to each level, than they already do.

References and further reading

Association of Language Testers in Europe (ALTE) *'Can-do' statements, http://www.alte.org/can_do/index.cfm*

Alderson, C (1989): *Bands and scores*, paper presented at the IATEFL Language Testing Symposium, Bournemouth, 17–19 November 1989.

Bachman, L (2001): Modern language testing at the turn of the century: assuring that what we count counts, *Language Testing*, 17 (1), 1–42.

Bachman, L F, Davidson, Ryan, K and Inn-Chull Choi (1995): *An investigation into the comparability of two tests of English as a foreign language*, Cambridge: Cambridge University Press.

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

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

Hamp-Lyons, L (1990): Second language writing: assessment issues, in Kroll, B (Ed.), *Second language writing assessment issues and options*. New York: Macmillan.

Hamp-Lyons, L and Prochnow, S (1991): The Difficulties of Difficulty: Prompts in Writing Assessment, in Anivan, S (Ed.) *Current Developments in Language Testing*, RELC Anthology Series 23, RELC.

Hamp-Lyons, L (1995): *Summary Report on Writing Meta-Scale Project*, UCLES Internal Report.

Linacre, J M (1988): FACETS: *A computer program for the analysis of multi-faceted data*, Chicago: MESA Press.

Messick, S A (1989): Validity, in Linn, R L (Ed.), *Educational Measurement* (3rd ed), NY: Macmillan, pp.13–103.

Morrow, K (1979): Communicative Language testing: revolution or evolution? In Brumfit, C and Johnson, K (Eds.) *The communicative approach to language teaching*, Oxford: Oxford University Press.

Papajohn, D (1999): The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants, *Language Testing,* 16(1) 52–81.

Pollitt, A (1991): Giving Students a Sporting Chance: Assessment by Counting and by Judging, in Alderson C & North, B (Eds.) *Language Testing in the 1990s*, London: British Council/Macmillan.

Read, J (1990): Providing Relevant Content in an EAP Writing Test, *English for Specific Purposes*, 9, 109–121.

Spolsky, B (1995): *Measured words*. Oxford: Oxford University Press.

Weir, C (1998): *Evaluation: Methods for Studying Programs and Policies.* New Jersey: Prentice-Hall Inc.

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

Insights into the FCE Speaking Test

YANG LU, PHD STUDENT, UNIVERSITY OF READING

Introduction

This article reports on a small-scale study of speaking performance at FCE level conducted at the University of Reading in Spring 2001. The study constituted a pilot for a larger scale investigation into the nature and assessment of discourse competence.

Research context

There has been growing interest in the validation of the notion of discourse competence, first proposed by Canale and Swain (1980) and later reformulated in Bachman's (1990) model of Communicative Language Ability. Results of previous studies investigating the validity of models of communicative competence have been mixed or contradictory: some suggest that grammatical and discourse competence are closely associated with each other (Bachman and Palmer 1982, Milanovic 1988), while others have suggested that grammatical competence and discourse competence are distinctive from each other (Allen et al. 1983).

This research study hypothesises that discourse competence is an independent component in test-takers' spoken communicative competence and that discourse performance correlates positively with other measures of test-taker performance.

The raw data for the study is a set of Speaking Test recordings for the First Certificate in English (FCE) examination, together with score data for the candidates. FCE Speaking Tests use a paired candidate format and candidate performance is assessed by two examiners: the Interlocutor, who manages the test interaction, and the Assessor, who observes the interaction. The Interlocutor awards each candidate one global score on a 5-point scale; the Assessor awards each candidate four analytical scores, also on a 5-point scale, for the following criteria: Grammar/Vocabulary, Pronunciation, Interactive Communication, and Discourse Management. (For more information, see the FCE Handbook, p.47).

For the purposes of this study:

- the two analytical scores – Grammar/Vocabulary and Pronunciation – were summed to produce an Analytic Linguistic Total (ALT) for each candidate;
- the two analytical scores – Interactive Communication and Discourse Management – were summed to produce an Analytic Discourse Total (ADT) for each candidate;
- the Overall Score (OS) for each candidate was used (a weighted score out of 40);
- the Interlocutor's global score (GS) for each candidate was used.

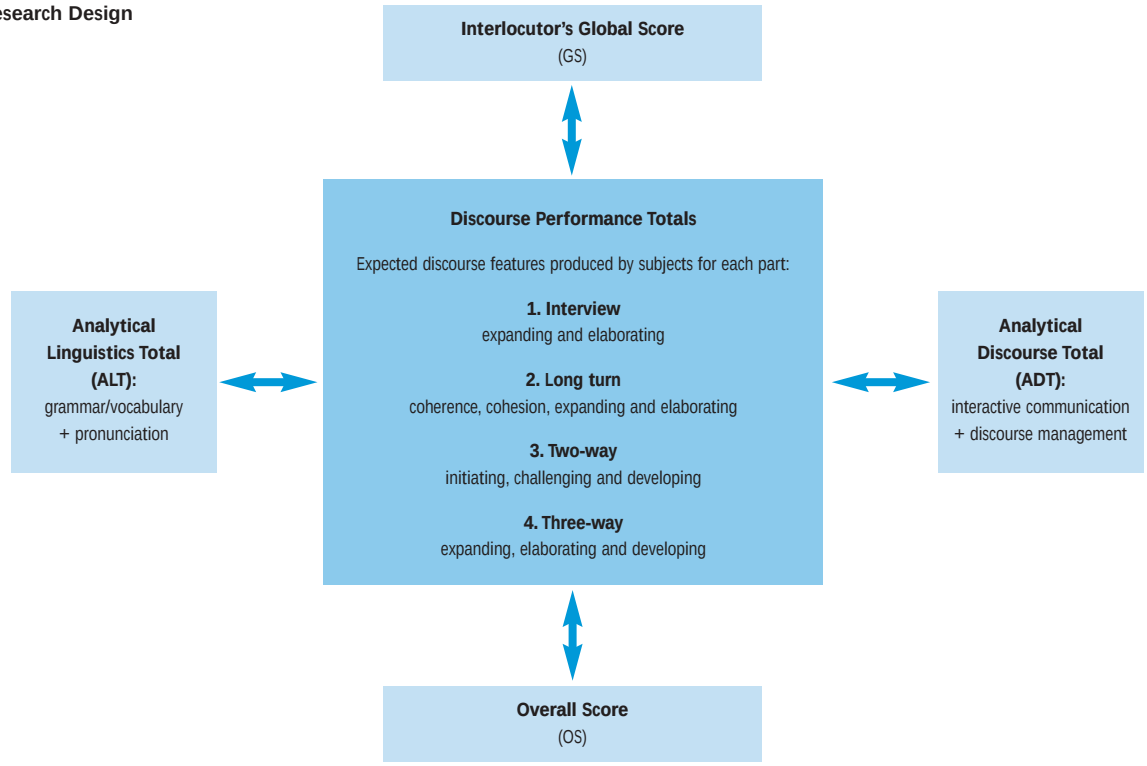
The rationale for grouping the analytical scores into two pairings (ALT and ADT) is that they seem to relate to two distinct competencies: ALT represents oral proficiency in terms of accuracy and appropriacy in using syntactic forms, lexical items and phonological features in English; ADT represents the ability to express ideas in coherent speech and to interact and develop appropriately in discourse. These two aspects of competence represent the candidates' oral linguistic and discourse competence which is to be investigated in relation to another measure – the Discourse Performance Total (DPT); DPT is a Discourse Analytical measure proposed by the research. The purpose of the pilot study was two-fold: to explore the notion of 'discourse competence' and to establish whether a discourse analytical approach offers a valid and reliable methodology for the larger scale research.

The specific hypotheses are that the test-takers' discourse performance, as measured using spoken discourse analysis methodology, will correlate positively with:

- their linguistic proficiency score (as represented by ALT);
- their discourse proficiency score (as represented by ADT);
- their overall oral proficiency score (as represented by OS);
- and the interlocutor's holistic judgement of their performance (as represented by GS).

To test the hypotheses, a research design was established to compare the four scores listed above (represented by the small, outer boxes in Figure 1 overleaf) with the results from discourse analysis of the FCE Speaking Test transcripts (represented by the central box in Figure 1); the aim is to see whether there are significant correlations (as shown by the double headed arrows).

Figure 1: Research Design



Methodology

Audio recordings of three live FCE Speaking Tests together with selected candidate information and score data were used for the study (see Table 1). All candidates took the Speaking Test in their home countries: Italy, Sweden and P. R. China. The test consists of four parts: Interview, Individual long turn, Two-way collaborative task, and Three-way discussion. The seven candidates used in this study took the test in two pairs and one group of three; Subjects 1, 2 and 3 in the data are a group of three, while the other four subjects form two pairs.

Using the Birmingham School Spoken Discourse Analysis Model (Sinclair and Coulthard, 1975, 1992), a Discourse Analysis framework was developed in relation to the FCE Speaking Test format. The framework lists the discourse features recognised as evidence of higher or high-level oral proficiency (Carroll 1980, Hasselgren 1996, Hoey 1991, Hughes 1989, McCarthy and Carter 1994, Shohamy 1994, Young 1995). The three recorded FCE Speaking Tests were then transcribed and analysed using a modified Birmingham School Spoken Discourse Analysis Model. Two important reasons for choosing the Birmingham School Model are:

1. It is a hierarchical structural-functional model that examines spoken language in ranks so that test-takers' oral production can be investigated at different discourse levels (Shohamy et al. 1986);
2. It describes not only the discourse functions of individual utterances but also how they combine to form larger discoursal units (Ellis 1985) and 'the sequencing of turns in conversations in terms of a set of functional "slots"' (Eggins and Slade 1997), so that learners' discourse can be examined in terms of its unique discoursal characteristics and patterns.

Table 1: Score data based upon subjects' FCE Speaking Test scores

Pair or Group	Subject	Analytical Linguistic Total (ALT)	Analytical Discourse Total (ADT)	Global Score (GS)	Overall FCE Speaking Test Score (OS) (a weighted score out of 40)
Group of Three	1	8	8	3	32.00
	2	9	8	4	36.00
	3	8	6	3	28.00
Pair 1	4	5	6	3	22.70
	5	7	7.5	4	30.00
Pair 2	6	8	9	4.5	34.70
	7	10	9.5	4.5	38.00

Table 2 shows in more detail the application of the Discourse Analysis framework to the recorded FCE Speaking Tests – with the categories for analysis and actual examples from the transcript data.

Table 2: Framework for Discourse Analysis

FCE Task	Categories for Analysis	Glossary	Examples from Data *(Bold italic text)*
Part 1: Interview	1. Expanding in responding	Acts that add information to an immediately prior move	- *What subjects did you least enjoy?* - *When I was at school that would be mathematics.* ***I was awful in the subject*** (elaborating). ***Several times I just give my sheets completely blank***. (expanding).
	2. Elaborating in responding	Acts that clarify, restate, exemplify an immediately prior move	
Part 2: Individual Long-turn	1. Coherence: a. opening & closing b. comparing & contrasting c. telling opinion	a. Utterances that signal the beginning or ending of the long turn b. Statements that tell the similarities or differences of the two pictures c. Ideas and personal preference about the pictures	***The other picture shows building er from the outside with trees and some grass*** (comparing). *It's nice* ***there's a lot grass, there's very bright*** (contrasting). Utterances for closing the long turn are absent in the data. See below for opening and telling opinions.
	2. Cohesive ties	Conjunctions, demonstratives & comparatives	- ***That****'s* (demonstrative) *not my style.* See below for conjunctions and comparatives.
	3. Lexical ties	Repetitions & synonyms	***OK*** (opening), ***so*** (conjunction) *we can see a students that is like* ***learning****, bored and struggling about what he is* ***studying*** (synonym), ***the subject*** (elaborating) ***from the other side*** (comparing) *a* ***man*** *in the coolest way er is thinking about his problems, he, I don't know,* ***mathematics or physics problem*** (elaborating) *in front of his computer* ***and*** (conjunction) ***in my opinion*** (telling opinion) *the situation of the* ***kids*** *for studying probably young* ***children*** (synonym) *in grammar school er it's* ***easier*** (comparative), ***but*** (conjunction) *involves much* ***more his emotional side than*** (comparative) *the* ***man*** (repetition) *in front of his* ***job***, ***everyday job*** (repetition and elaboration).
	4. Elaborating & expanding when describing	Same as in Part 1, but they are used in describing the pictures	*I guess, there is more about myself I think.* ***It's much more interesting*** (expanding) See above for elaborating acts.
Part 3: Two-way Collaborative task	1. Initiating	Utterances that elicit responses	- ***So you like taking pictures*** (initiating) - *Yes, taking pictures, it's important*
	2. Developing	Utterances that develop one's own or other's ideas	- *There is a supermarket er on the first floor. I think people live there will be easy to buy thing.* - ***And feel more comfortable, because there is communc, communca, so er*** (developing)
	3. Challenging	Utterances that contradict previous move toward further discussion	- *Because I like I like private house, instead being disturbed by others, we can do our own things* - *Yeah*, - ***I don't agree with you*** (challenging) - *More nature, more fresh air*
Part 4: Three-way discussion	1. Expanding	Same as in Part 1	See examples in Part 1
	2. Elaborating	Same as in Part 1	See examples in Part 1
	3. Developing	Same as in Part 3	See examples in Part 3

Table 3: Scores for Subjects' Discourse Performance by Discourse Analysis

Group or Pair	Subject	Part 1: Interview	Part 2: Individual long turn	Part 3: Two-way collaborative task	Part 4: Three-way discussion	Discourse Performance Totals
Group of Three	1	2	17	12	7	**38**
	2	9	17	11	4	**41**
	3	2	11	8	7	**28**
Pair 1	4	5	17	8	12	**42**
	5	4	15	11	11	**41**
Pair 2	6	6	17	14	11	**48**
	7	13	17	13	8	**51**

After identifying these discourse features, the number of features in different categories were summed to produce a Discourse Performance Total (DPT) for each subject (see Table 3).

Minitab statistical software was used to correlate the DPT scores shown in Table 3 with the scores given in Table 1.

Results and discussion

As Table 4 shows, the DPT scores derived from the application of discourse analysis to the FCE Speaking Test transcript data correlated positively and significantly with the subjects' ADT and GS scores. The DPT scores correlated positively but with no significance with the ALT and OS scores.

Table 4: Correlations of DPT scores with other measures

Categories of Correlations	Correlation Coefficients	P-Value (0.05)
DPT & ALT	0.248	0.592 not significant
DPT & ADT	0.779	0.039 significant
DPT & GS	0.777	0.040 significant
DPT & OS	0.527	0.225 not significant

These results partly support the hypotheses outlined above and suggest that though the Assessors, Interlocutors and the discourse analyst were adopting varying approaches to assessing candidates' spoken discourse competence, they tend to agree with each other in most cases.

A positive and significant result is not really surprising in the case of DPT and ADT since they examine similar discourse features such as expanding in responding, coherence in long turn and initiating in conversations; such a result was more surprising in the case of the GS which is based on a Holistic Rating Scale consisting of mixed criteria (accuracy of grammar and vocabulary, coherence, comprehensibility, success of communication). If these results were to be repeated with a larger number of subjects, it might imply that spoken discourse competence, as well as linguistic competence, is being captured within the Interlocutor's global score. Is this because an Interlocutor actually interacts with the candidates and is therefore equally sensitive to their discoursal behaviour? Or is discourse performance predictive of test-takers' syntactic and lexical proficiency levels? Or, were the Interlocutors ignoring primarily linguistic criteria and focusing on discoursal competence?

DPT correlated positively but not significantly with ALT and OS. This may be because combining scores for candidates' grammatical, lexical and phonological performance is not representative of the test-takers' discoursal performance; they are the results of measuring different aspects of test-takers' oral proficiency. It may also imply that the Assessors' judgement of the test-taker's linguistic performance is little related to the results of the discourse analysis, possibly because the former are mostly concerned about grammatical, lexical and phonological accuracy and appropriacy.

In any event, the candidates' discourse competence does appear to constitute an independent competency among the different components of their overall communicative competence in spoken language. There seems to be a distinction between spoken *discourse* competence and spoken *linguistic* competence or the overall spoken communicative competence. They may sometimes cluster together as the significant positive correlation between DPT and GS shows.

Conclusion

The results of this small-scale pilot study indicate that in the case of the discourse-related criteria and the interlocutors' holistic judgment of test-takers performance there were significant correlations with the measure of discoursal features adopted here. This is seen as a positive result in that it supports the hypothesized (on the part of the test developer) relationship between the two. It can also be seen as offering evidence of the validity of the separate discourse management scale used in the test. The fact that there was no significant relationship between the discoursal features and the FCE linguistic measures appears to confirm this evidence.

The pilot study has shown that the proposed PhD research is feasible in terms of appropriacy of data and the test-takers, practicality of the research design, the statistical procedure, and the workability of the discourse analysis approach. It should be remembered, however, that the findings from the pilot study are based on a relatively small amount of data; the main PhD study will apply the techniques described above to a much larger sample of over 60 FCE Speaking test transcripts and the results will be reported in a future *Research Notes*.

References and further reading

Allen, P, Cummins, J, Mougeon, R and Swain, M (1983): *The Development of Bilingual Proficiency: Second Year Report and Appendices*, Toronto, Ontario: Modern Language Centre, Ontario Institute for Studies in Education.

Bachman, L and Palmer, A (1982): The construct validation of some components of communicative proficiency, *TESOL Quarterly* 16(4), 449–465.

Bachman, L (1990): Constructing measures and measuring constructs, in Harley et al (1990a) pp.26–38.

Canal, M and Swain, M (1980): Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing, *Applied Linguistics* 1 (1), 1–47.

Carroll, B J (1980): *Testing Communicative Performance,* Oxford: Pergamon.

Eggins, S and Slade, D (1997): *Analysing Casual Conversation*, London and New York: Continuum.

Ellis, R (1994): *The Study of Second Language Acquisition,* Oxford: Oxford University Press.

Harley, B, Allen, P, Cummins, J and Swain, M (Eds.)(1990a): *The development of second language proficiency*, Cambridge: Cambridge University Press.

Harley, B, Allen, P, Cummins, J and Swain, M (1990b): The nature of Language Proficiency, in Harley et al, pp.7–25.

Hasselgren, A (1997): Oral test subskill scores: what they tell us about raters and pupils, in Huhta, A, Kohonen, V, Kurki-Suonio, L & Luoma, S (Eds.) *Current Developments and Alternatives in Language Assessment – Proceedings of LTRC 96*, Rowley, Mass: Newbury House, pp.241–256.

Hoey, M (1991): Some Properties of Spoken Discourse, in Bowers, R and Brumfit, C (Eds.) *Applied Linguistics and English Language Teaching*, Basingstoke: Macmillan, pp.65–84.

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

Milanovic, M (1988): *The construction and validation of a performance-based battery of English language progress tests*, Unpublished PhD Dissertation, University of London.

McCarthy, M and Carter, R (1994): *Language as Discourse: Perspectives for Language Teaching*, London and New York: Longman.

Shohamy, E, Reves, T and Bejerano, Y (1986): Introducing a New Comprehensive Test of Oral Proficiency, *ELT Journal* 40, 212–220.

Sinclair, J and Coulthard, M (1975): *Towards an Analysis of Discourse*, Oxford: Oxford University Press.

Sinclair, J and Coulthard, M (1992): Towards an Analysis of Discourse, in Coulthard, M (Ed.) *Advances in Spoken Discourse Analysis*, London and New York: Routledge, pp.1–34.

Swain, M (1985): Large-scale communicative language testing: A case study, in Lee, Y P, Fork, A C Y, Lord, R and Low, G (Eds.) *New Directions in Language Testing*, Oxford: Pergamon, pp.35–46.

Young, R (1995): Conversational Styles in Language Proficiency Interviews. *Language Learning* 45 (1), 3–42.

Conference Announcements

Language Testing Forum 2003

This will be hosted by Cambridge ESOL in November 2003. Further details will be available shortly on our website and will be announced in the next issue of Research Notes (May 2003).

25th Language Testing Research Colloquium

The next Language Testing Research Colloquium will take place between July 22nd–25th 2003 at Reading University, hosted by the Testing and Evaluation Research Unit. The conference theme will be *Learner Interfaces with Language Testing & Assessment*. There will be a range of research papers, symposia, poster sessions and a research networking session. The organisers are Barry O'Sullivan (The University of Reading), Pauline Rea-Dickins (The University of Bristol) and Jayanti Banerjee (Lancaster University).

Full details are available from the conference website: http://www.rdg.ac.uk/AcaDepts/ll/teru/ltrc2003/homepage.htm

AAAL and TESOL 2003

The annual conferences of the American Association for Applied Linguistics (AAAL) and Teaching English to Speakers of Other Languages (TESOL) take place in March 2003 in the USA. As usual, the two conferences will run back to back: March 22–25 (AAAL) and March 25–29 (TESOL). AAAL will be held in Arlington, Virginia, and TESOL will follow on in Baltimore, Maryland. For more details, visit the conference websites at:

AAAL: http://www.aaal.org/aaal2003/index.html
TESOL: http://www.tesol.org/conv/t2003/pp/99-sitemap.html

Research and Validation staff will be presenting at both events and there will be a fully-equipped stand at the TESOL conference carrying a wide range of Cambridge ESOL publications.

Converting an Observation Checklist for use with the IELTS Speaking Test

LINDSAY BROOKS, PHD STUDENT, OISE (THE ONTARIO INSTITUTE FOR STUDIES IN EDUCATION), UNIVERSITY OF TORONTO

Introduction

An observation checklist (OC) was originally developed (see O'Sullivan, Weir & Saville, 2002; Saville & O'Sullivan, 2000) for analysing the functions elicited in the paired speaking tasks of the Main Suite examinations. The OC consists of an extensive table of informational, interactional and managing interaction functions or operations that tend to be elicited in the language tasks of the Main Suite speaking tests such as the FCE and CPE. Designed to evaluate task output in 'real time' (see Saville, 2000), the checklist provides an efficient means of investigating the variation in language elicited in different task types; it identifies the language functions associated with particular tasks and is capable of producing a profile of the language elicited across several tasks within a speaking test. This information about the language elicited in the speaking tests provides important validation evidence since the predicted versus actual language elicited in each task can be compared, task comparability between tests can be checked, and both of these in turn can inform task design. Although the checklist was originally designed for use in the revision of the speaking tasks for the CPE, it seemed like it might be possible to extend the use of the OC to the revised IELTS Speaking Test, taking into account its singleton format and the different task types of the test. The focus of this article then is on the investigation and subsequent conversion of the OC for use with the speaking tasks in the IELTS Speaking Test.

The Revision of the Observation Checklist

To investigate the feasibility of using a checklist with the IELTS Speaking Test, the OC was applied to seven videotaped IELTS speaking tests from the training/certification video developed for the introduction of the revised IELTS test which occurred in July 2001 (see Taylor, 2001). The OC and the training materials (see O'Sullivan, 2001) had been developed using videotaped FCE tests but because IELTS tests are always audiotaped rather than videotaped, the possibility of applying the checklist to audiotaped tests needed to be examined. Therefore, the OC was also applied to nine audiotaped IELTS tests and any problems with applying the OC to this format were noted. In the audiotaped mode, despite not being able to see the interaction between each candidate and examiner, no difficulty was experienced in applying the checklist and listening for the operations/functions on the OC. An advantage of the audiotaped format over the videotaped format was that it was possible to concentrate on the language elicited with no distractions from the videotape. The only limitation of using the OC in the audiotaped format was that detection of non-verbal communication was not possible.

Audio- vs. Videotaped Tests

Although the OC seemed to be suitable for the audiotaped format, it was necessary to determine if the functions "observed" were the same as those observed in the videotaped format. Therefore, several months later, allowing for sufficient time to pass for the investigator to have forgotten the initial results, the OC was applied to the same seven IELTS speaking tests that were on the training videotape; however, for this second application of the OC, the IELTS speaking tests had been transferred to an audiotape. A comparison of the results of using the OC on the audiotaped and the videotaped tests provides insight into 1) intra-rater reliability of applying the OC to the same speaking tests in the different formats and 2) any differences resulting from the two modes of delivery and therefore the practicality of applying the OC to audiotaped tests, which are more readily obtainable than videotaped tests.

A comparison of the observed checklist operations in the audiotaped and videotaped applications resulted in a high degree of reliability. Any discrepancies between the two formats involved under-representation of operations in the videotaped format, possibly due to the added complexity of watching the video and scanning the OC for the functions. All of the 144 observed operations in the seven videotaped tests were also "observed" or more accurately heard in the audiotaped format. An additional 23 functions, which were not observed in the videotaped format, were noted in the audiotaped format. This represents an overall 86% agreement between the audiotaped and videotaped applications of the OC. The intra-rater agreement for each of the seven speaking tests (audiotaped versus videotaped versions) ranged from 77% to 95% (see Table 1 for a breakdown of the reliability for each of the seven tests).

Table 1: Reliability of applying the OC to the audiotaped and videotaped IELTS tests

Candidate	Functions Agreeing	Total Functions*	Reliability (%)
1	24	28	86
2	19	21	90
3	21	26	81
4	17	22	77
5	21	23	91
6	22	26	85
7	20	21	95
	144	167	86

*this represents the total number of functions as "observed" when applying the OC to the audiotaped IELTS tests

Tailoring the checklist to IELTS

Once the practicality and comparability of using an observation checklist in the context of audiotaped IELTS Speaking Tests had been determined, in consultation with the original developers of the OC, further changes were made to the checklist to convert it for use with the IELTS Speaking Test. References to non-verbal communication were removed and references to a partner were changed to reflect the candidate's interaction with an examiner rather than another candidate. Other language in the checklist was modified slightly to fit the output of the tasks of the IELTS Speaking Test. The training materials developed for the original OC include examples of all of the operations on the checklist taken from other tests in the UCLES Main Suite so examples from IELTS tests were included to make the training materials specific to applying the checklist to this test.

Conclusion and Future Directions

Although the checklist was initially designed for use with the CPE, the results of this study show that the OC is a flexible instrument that can be adapted for use in other tests, such as the IELTS Speaking Test. This is potentially significant in that if in the future the intention is to conduct a larger scale study of the IELTS speaking test for the purposes of test validation evidence, it will be possible to use the audiotaped tests that are routinely collected by the test centres as part of the testing process. Because the small trial of audiotaped versus videotaped IELTS Speaking Tests was conducted using the original OC, future work on the use of the checklist with IELTS will include replicating this small trial with the revised IELTS OC. Obtaining an inter-rater reliability estimate by having another investigator go through the process of applying the OC to the seven speaking tests in both the audiotaped and videotaped versions would also provide insights into how well and accurately the OC can be used to provide an overview of the functions and operations observed in the IELTS Speaking Test.

References and further reading

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

Saville, N (2000): Using observation checklists to validate speaking-test tasks, *Research Notes 2*, 16–17.

Saville, N and O'Sullivan, B (2000): Developing observation checklists for speaking tests, *Research Notes 3*, 6–10.

Taylor, L (2001): Revising the IELTS speaking test: Developments in test format and task design, *Research Notes 5*, 2–5.

Conference Reports

The last few months have seen a number of key language testing and other conferences which have been attended by Research and Validation staff. Three conferences which we contribute extensively to are reported below, namely the Language Testing and Research Colloquium, the Language Testing Forum and the British Association for Applied Linguistics conferences.

A list of all of the conferences that Cambridge ESOL will attend in the coming months can be found at: http://www.cambridge-efl.org/confex/index.cfm

Language Testing Research Colloquium 2002

The 2002 Language Testing Research Colloquium was held in Hong Kong from 12 to 15 December. The overall theme was 'Language Testing in Global Contexts' – a theme of particular relevance to Cambridge ESOL which now administers language tests in over 150 countries of the world. The work of Cambridge ESOL's Research and Validation Group was once again well represented at this key international conference for the world-wide language testing community.

In a paper, entitled *Plurilingualism and partial competence: implications for language assessment*, Nick Saville outlined the related concepts of plurilingualism and partial competence as defined by the Council of Europe's Common European Framework of Reference for Languages. He looked at the implications for assessment, and especially for education authorities and assessment agencies.

Lynda Taylor presented a paper (written jointly with Stuart Shaw) entitled *Revising assessment criteria and scales: insights from the rater community* in which she reported on a survey carried out with writing examiners in the context of the IELTS Writing Revision Project. The paper showed how the rich insights offered by the 'community of raters' can be instrumental in helping to develop writing assessment criteria and scales.

Professor Alan Davies chaired a symposium which focused on the theme of *Shifting conceptions of reliability* and featured three separate contributions from Ardeshir Geranpayeh (with Neil Jones), Nick Saville (with Neil Jones), and Lynda Taylor. Descriptions of levels of language ability, such as the Common European Framework, offer language testers the chance to give exams more useful meaning, relating to well-understood descriptions of communicative language ability. But linking exams into a larger interpretative framework also calls into question the traditional concept of reliability as an index of test quality. The symposium reviewed the usefulness of traditional measures of reliability and considered alternative, complementary approaches to determining test quality. In terms of providing evidence of reliability, as one of the essential features of a test, three areas were touched on by the seminar:

1. The potential a testing system has for being reliable (as a result of its design features);
2. The estimation of reliability based on data and analysis of the test when it is administered;
3. The reporting and explanation of the reliability estimates to users of the test, particularly the need to frame an explanation within an overall picture of the test's qualities.

Key themes or strands which seemed to emerge in the other papers presented included: technological applications in language testing; issues in standards-based language assessment; the social context and implications of language testing; the role of cross-cultural factors; and the processes and products involved in writing assessment. Interestingly, these are all issues which we report on regularly in *Research Notes*.

The quality of the conference papers was generally very good, and the poster and research network sessions enabled valuable face-to-face, more informal interaction between individual researchers and others who were interested or working in similar fields. It was especially pleasing to see so many language testers present at the conference from the countries of East and South Asia.

As always the conference proved to be an enjoyable and highly stimulating gathering of language testers from all corners of the world. Despite often working in very differing political, social, and educational measurement contexts, we discovered much of common interest and came away with a better grasp of some of the specific issues that are challenging our language testing colleagues in other parts of the world.

Language Testing Forum 2002

The Language Testing Forum 2002 was hosted by the Testing and Evaluation Research Unit at the University of Reading in November. A number of presentations were made by members of the Cambridge ESOL Research and Validation Unit. Despite last minute changes to the arrangements because of a fire fighters' strike, the conference proved very successful, and delegates now look forward with enthusiasm to the 2003 LTF to be held in Cambridge.

A theme of the conference was benchmarking and the development of common frameworks for language tests, and the topic of benchmarks was the basis for a lively panel discussion between Barry O'Sullivan (the University of Reading), Nick Saville (Cambridge ESOL) and Charles Alderson (Lancaster University). The participants discussed issues surrounding benchmarking, emphasising the need for a quality control dimension in relating tests to common frames of reference such as the ALTE levels and the Common European Framework. In a related paper, Neus Figueras, from the Departament d'Ensenyament Generalitat de Catalunya in Barcelona, described some of the challenges involved in relating locally developed test instruments to the Common European Framework (CEF).

Miyoko Kobayashi (Kanda University of International Studies) opened the conference. She reported on research carried out among Japanese university students into the characteristics of short answer questions for reading comprehension. She outlined a principled approach to reading test construction on the basis of a model of text and item characteristics and suggested promising avenues for further investigation of reading comprehension and its component abilities.

Brian Richards (University of Reading) described the writing tasks given to UK children at Key Stages 1, 2 and 3 and demonstrated convincingly that there would be problems in using totally automated essay assessment. Neil Jones (Cambridge ESOL) talked about validating the writing component of CELS (the Certificates in English Language Skills). Jay Banerjee and Dianne Wall from Lancaster University talked about reporting the progress of students at the end of the Lancaster University four-week pre-sessional English course. Ardeshir Geranpayeh (Cambridge ESOL) reported on some of the validation activities carried out during the construction and validation of a German placement test that Cambridge ESOL and the Goethe Institute are producing and which is to be given on computer or in a paper and pencil version, and Pavlos Pavlou, University of Cyprus, talked about test takers' rights and possible violations to these rights.

The conference closed with a talk by Vita Kalnberzina (Lancaster University) who used Structural Equation Modelling as one of her tools for assessing the effect of test anxiety on students taking the final year English test in Latvian schools. She based her research on the Bachman and Palmer model of language proficiency, and found that although test anxiety as such did not appear to have any great effect on students' test performance, affective schemata, strategic competence and language performance were clearly linked.

There were eight paper presentations altogether and a poster session at which students shared their developing research plans. For a complete list of the conference papers and posters, see *Language Testing Update*, Issue 32.

British Association for Applied Linguistics Conference 2002

The thirty-fifth BAAL conference took place over three days in Cardiff in September. The conference theme was 'Applied Linguistics and Communities of Practice' which was chosen to 'broadly reflect the various communities of practice where applied linguistic research has been relevant over the last few decades and also poses new challenges' (BAAL website).

Various communities of practice were described and discussed during this conference, from the community of raters (one of Cambridge ESOL's contributions) to the more typical institutional communities such as healthcare, global education, the legal system and so forth.

Professor David Crystal (University of Wales) in the opening plenary expressed a need for applied linguists to work in specific areas including the theatre and internet communication, some of the 'Final Frontiers in Applied Linguistics'. These and other areas

challenge the boundaries of Applied Linguistics and would, Crystal hoped, receive coverage at the conference. Crystal also suggested that cultural understanding leads to a grasp of English and vice versa although he warned that linguists should be wary of imposing cultural imperialism on learners, something of particular relevance to the work of Cambridge ESOL.

The other plenary speakers were Celia Roberts (King's College, London) and John Swales (University of Michigan). Roberts spoke about 'Applied Linguistics Applied' in which she focussed on the 'applied' part in a reflexive way, using a case study of medical research to show how Applied Linguists and medical specialists can misunderstand one another's opinions, even where both are aiming at the same outcome. John Swales considered whether a university is a community of practice and used evidence from the Michigan Corpus of Academic English (MICASE) to explore whether different types of oral communication are university wide or differ according to department.

Cambridge ESOL contributed papers on using wordlists in language testing (Fiona Ball) and the assessment of pen-and-paper and computer based IELTS academic writing tasks (Russell Whitehead) together with a plenary on making judgements about language using insights from the 'community of raters' (Lynda Taylor, Stuart Shaw and Alan Tonkyn, University of Reading). There were also contributions on health communication, interpreting and translating, native speaker norms, vocabulary acquisition and teaching and testing in various contexts, all of which are of relevance to Cambridge ESOL.

Although the number of language testers per se was relatively small at this conference, Cambridge ESOL understands the importance of keeping up-to-date with current theories and practice in the field of Applied Linguistics. We look forward to the next BAAL conference which will take place in Leeds in September 2003.

For further information see the organisation's website:

IELTS MA Dissertation Award 2003

As part of the tenth anniversary of IELTS in 1999, the IELTS partners – University of Cambridge ESOL Examinations, The British Council, and IDP Education Australia – agreed to sponsor an annual award of £1000 for the MA dissertation in English which makes the most significant contribution to the field of language testing. In its inaugural year the award went to joint winners in Australia and Canada. For 2001 the award went to a Korean student studying at the University of California, Los Angeles (UCLA) (see page 24). The IELTS Research Committee, which comprises representatives of the three partner organisations, decided that for 2002 no award would be made.

For 2003, the entry procedures and timetable for the award are as follows:

Submission and evaluation procedures

Dissertations will only be considered eligible if they were submitted and approved by your university in 2002. Dissertations completed in 2003 will not be considered eligible for the 2003 award but may be submitted the following year.

Submissions should be for dissertations written in partial or total fulfilment of the requirements for an MA degree or its equivalent.

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

Dr Lynda Taylor / Stuart Shaw
University of Cambridge ESOL Examinations
1 Hills Road
Cambridge, CB1 2EU
United Kingdom

- The IELTS Research Committee will review the submissions and shortlist potential award winners;
- For all shortlisted dissertations a full copy of the dissertation will be requested and a further reference may be sought;
- Shortlisted dissertations will be reviewed and evaluated by the IELTS Research Committee according to the following criteria:
  - Rationale for the research;
  - Contextualisation within the literature;
  - Feasibility of outcomes;
  - Design of research question(s);
  - Choice and use of methodology;
  - Interpretation and conclusions;
  - Quality of presentation;
  - Use of references;
  - Contribution to the field;
  - Potential for future publication.
- The Committee's decision is final.

Timetable

The following timetable will apply in 2003:

- **1 June** Deadline for submission of dissertation extracts and supervisor's reference to University of Cambridge ESOL Examinations
- **1 August** Deadline for submission of full copies of shortlisted dissertations (and further references if required)
- **October/November** Meeting of IELTS Research Committee
- **November/December** Announcement of award

Details of the application process for the IELTS MA Dissertation

Award 2003 can also be found on the IELTS website – www.ielts.org.uk. Please note that submission details may change from year to year and it is therefore important that the most current procedures are consulted.

Presentation of 2001 IELTS MA Award at LTRC

At LTRC in December the 2001 winner, Sang Keun Shin, was presented with his IELTS MA award by Cambridge ESOL representatives Lynda Taylor and Nick Saville.

Nick Saville (Cambridge ESOL); Angel Lam (IELTS Manager/IDP, IELTS CEPA Program), Sang Keun Shin, Georgina Pearce (Deputy Director, English Language Centre, The British Council)

Other news

Use of Cambridge ESOL's Materials by Researchers

Cambridge ESOL regularly receives requests for data and materials from researchers and research students (details of the procedure to follow can be found in *Research Notes 4*; an updated version is available from contacts below).

We are also interested in hearing about other research that has been completed or is planned/underway that uses any of our publicly available materials (e.g. sample papers and answers from the website). We may, in some cases, be able to suggest or provide more appropriate or up-to-date materials for the study, or may request a copy of the research for our ESOL library.

For further information please contact: Lynda Taylor (taylor.l@ucles.org.uk) or Fiona Barker (barker.f@ucles.org.uk).

New Research and Validation Group Staff

We recently welcomed two new staff members to the Research and Validation Group which takes the number of permanent staff to 21.

Dr Caroline Clapham took up a new post as IELTS Validation Officer in September 2002. Her responsibilities include co-ordinating and carrying out research to underpin the IELTS examination. Caroline has worked at Lancaster University as researcher and lecturer since 1974, and has previously been connected with many IELTS-related projects. She is the author of Studies in Language Testing volume 4 'The Development of IELTS: a Study of the Effect of Background Knowledge on Reading Comprehension'.

Anthony Green joined Cambridge ESOL in October 2002 as Validation Officer/Grading Co-ordinator. Tony is currently completing a PhD at the University of Surrey supervised by Professor Cyril Weir concerned with washback from the IELTS Writing test on preparation for academic study in the UK. Previously, while still at Reading University, he was also involved in a number of IELTS-related research projects.

Accreditation of Cambridge ESOL Teaching Awards

The Cambridge ESOL Teaching Awards – including the well-established Certificate in English Language Teaching to Adults and Diploma in English Language Teaching to Adults (CELTA and DELTA) – have received official recognition in the UK from the Qualifications and Curriculum Authority (QCA). The CELTA award has been accredited by the QCA as the Cambridge ESOL Level 4 Certificate in Teaching English to Speakers of Other Languages (CELTA), whilst the QCA title for DELTA is the Cambridge ESOL Level 5 Diploma in Teaching English to Speakers of Other Languages (DELTA). These awards are now more valuable for language teachers and their accreditation reinforces the prestige of these internationally recognised qualifications.

Grade Statistics and other Support Materials

Grade statistics for many Cambridge ESOL examinations are available from the support pages on our website: http://www.cambridge-efl.org/support/

The statistics show the percentage of candidates by country gaining each grade of a particular examination. The most recent statistics are for 2001 – those for 2002 will be posted later this year.

The support pages also include:

- past papers
- handbooks, reports and regulations
- newsletters (*Cambridge First* and *Research Notes*)
- a list of publishers that produce learning materials for our examinations
- special arrangements information
- a list of seminars for teachers
- details of open centres

New ESOL CentreNet Service

ESOL CentreNet is an on-line service exclusively for authorised Cambridge ESOL examination centres. It provides on-line tools to help centres administer Cambridge ESOL exams. Through ESOL CentreNET centres can send Cambridge ESOL entries, receive results and much more. The site offers Administration and Professional Support in various areas including exams processing, obtaining examination and publicity materials, information about exam-related events, software and news to keep centres up-to-date with administrative and personnel changes in Cambridge.