5

ResearchNotes

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE
ENGLISH AS A FOREIGN LANGUAGE (EFL)

JULY 2001

EFL Information
University of Cambridge
Local Examinations Syndicate
1 Hills Road
Cambridge CB1 2EU
United Kingdom

Tel: +44 1223 553355
Fax: +44 1223 460278
e-mail: efl@ucles.org.uk
www.cambridge-efl.org

University of Cambridge
Local Examinations Syndicate

ALTE Association of Language Testers in Europe

EAQUALS ASSOCIATE MEMBER

ResearchNotes

Introduction

Welcome to the fifth issue of *Research Notes*, the UCLES EFL newsletter about current developments in research, test development and validation issues.

July 2001 sees the introduction of the revised IELTS Speaking Test. Following on from her recent article describing the development of revised assessment criteria and rating scales, Lynda Taylor explains the rationale for the new test format and reports on some of the research and validation work which underpins it.

Contents

In *Research Notes 2* Neil Jones introduced the ALTE Can Do Project which aims to provide a comprehensive description of what language users can typically do with the language at each level, in the various language skills and in a range of contexts. This issue provides an update on progress, reporting on the calibration of the individual Can Do statements on the basis of empirical data from self-report questionnaires.

Direct tests of written language performance have long been considered important when profiling a learner's communicative competence. But what are the key features which distinguish levels of performance among L2 writers, and how can these help to provide a descriptive scale of L2 writing proficiency? Roger Hawkey reports on Phase 2 of our ongoing Common Scale for Writing Project.

The stability of test results over time is a major concern of test designers, especially in relation to computer-based testing. Ardeshir Geranpayeh follows up an earlier study by Neil Jones (reported in *Research Notes 3*) to examine the reliability of computer-based BULATS by means of the test-retest method.

Nick Saville previews some forthcoming volumes in the Studies in Language Testing series which will report on various UCLES test revision projects; and Lynda Taylor highlights plans over the coming months to publish the first in a series of *EFL Research Papers*. We also report on the restructuring which has taken place within the UCLES Validation Group since we published our first issue of *Research Notes* in March 2000.

Research Notes is intended to reach a wide audience of people involved in Cambridge examinations around the world and also people who are interested in the theoretical and practical issues related to language assessment. We would be very interested to hear your views on the newsletter – whether you find it interesting and useful, how appropriate you find the level of presentation and if there are any topics you would like us to cover. You can e-mail research.notes@ucles.org.uk or write to us at the address on page 23.

Research Notes is delivered to all UCLES EFL centres and other key contacts. If you would like to receive additional copies or if you would like a personal subscription to the newsletter, please complete and return the form on page 23.

Published July 2001

Revising the IELTS Speaking Test: developments in test format and task design

Lynda Taylor, Senior Research and Validation Co-ordinator

Issue 4 of *Research Notes* (February 2001) reported on the project to revise the IELTS Speaking Test, in particular some of the development and validation work to revise the assessment criteria and rating scales. This follow-up article focuses on the rationale for the revised test format and the tasks which are included in it. (A brief background to the International English Language Testing System was given in *Research Notes 4* so will not be repeated here; further information on IELTS is available from www.ielts.org)

Background

The revision project for the IELTS Speaking Test began in early 1998 with identification of issues needing to be addressed. This was informed from a number of sources including: a review of the routinely collected candidate score and test performance data for the operational IELTS speaking test; a review of theoretical and empirical studies on the test conducted between 1992 and 1998 (e.g. Ingram and Wylie, 1993; Brown and Hill, 1998, Merrylees and McDowell, 1999); a review of other research into speaking assessment, together with work on speaking test design for the other Cambridge EFL tests (see Lazaraton, in press/2001). Consultation with a range of stakeholders confirmed that certain features of the existing speaking test should be retained - the 1-to-1 format (one candidate and one examiner, with audio recordings for checking and monitoring); the overall test length (max. 15 minutes); and the multi-phase approach.

The original IELTS Speaking Test was designed with five phases, with phases 2-4 designed to push the candidate progressively to his/her 'linguistic ceiling'. However, analyses of the operational use of the test indicated that Phases 3 and 4, in which the candidate was required to elicit information, to express precise meaning and attitudes, and to speculate, did not always elicit a 'richer' performance; moreover, these elicitation problems led, in turn, to variations in amounts and type of examiner-talk.

No change was envisaged to the underlying construct/s of spoken language proficiency. Current cognitive views of the speech production process (e.g. Levelt, 1989; Garman, 1990) suggest that the proficient L2 speaker will possess the following competence:

a) a wide repertoire of lexis and grammar to enable flexible, appropriate, precise construction of utterances in 'real time.' (The knowledge factor)

b) a set of established procedures for pronunciation and lexico-grammar, and a set of established 'chunks' of language, all of which will enable fluent performance with 'on-line' planning reduced to acceptable amounts and timing. (The processing factor)

In addition, spoken language production tends to be based in social interaction, to be purposeful and goal-oriented within a specific context, and, while it is capable of being routine and predictable, it also has the capacity for relative creativity and unpredictability.

Research in recent years has highlighted various features that are characteristic of more or less proficient oral performances (see Tonkyn and Wilson, forthcoming, for a list of useful studies which can help oral test designers identify theoretically relevant and helpfully discriminating features of performance).

The IELTS revision project set out therefore to develop a clearer specification of the speaking test, in terms of input and expected candidate output, and to revise the test format and tasks so as to elicit an appropriate sample of spoken language for assessment purposes. A further objective was to increase standardisation of test conduct by introducing an examiner frame.

Phase 1 : Consultation, initial planning and design (May-Dec 1998)

The Working Party began by producing a revised content specification, test format and sample tasks (see Table 1). The revised test format was divided into 3 parts, with each part designed to fulfil a specific function in terms of interaction pattern, task input and candidate output.

Table 1: Format of revised speaking test

Part	Nature of interaction	Timing
Part 1		
Introduction and interview	Examiner introduces him/herself and confirms candidate's identity. Examiner interviews candidate using verbal questions based on familiar topic frames (e.g. home/family, interests, etc.).	4-5 minutes
Part 2		
Individual long turn	Examiner asks candidate to speak for 1-2 minutes on a particular topic based on written input in the form of a general instruction and content-focused prompts. Examiner asks one or two rounding-off questions at the end of the long turn.	3-4 minutes (incl. 1 minute preparation time)
Part 3		
Two-way discussion	Examiner asks candidate to participate in discussion of more abstract nature, based on verbal questions thematically linked to Part 2 prompt.	4-5 minutes

The revised test format was designed to suit both Academic and General Training candidates; this was especially important given the substantial growth in the GT candidature over recent years. Part 1 deals with familiar topics that all those taking the test will be able to respond to. Part 2 replicates presentation skills required in academic seminars, but uses topics accessible to all. Part 3 invites the less competent speakers to explain and describe, while the more proficient speakers have the opportunity to develop arguments, justify opinions, analyse and speculate. The test has been designed so that there is a progression from familiar topics to more unfamiliar ones – a move from less to more challenging subject matter. However, candidates will now be assessed on a sustained performance over the 3 parts of the speaking test: they are no longer seen as moving towards a 'linguistic ceiling' as the test proceeds.

The long turn in Part 2, which provides the candidate with an opportunity for sustained language production and for taking the initiative in the interaction, is a particular and distinct enhancement to the current test. Part 2 also includes preparation time since studies by Wigglesworth (1997) and Skehan and Foster (1997) indicate that preparation time and forethought can enhance performance. The candidate's prompt card for Part 2 provides a context and content points to guide them; these do not need to be followed strictly, but they give valuable support to weaker candidates. The prolonged turn also temporarily frees the examiner from the active interlocutor role and allows him/her time to focus entirely on performance. This is an important factor in strengthening the reliability and accuracy of the assessment.

A significant change in the test procedures was the introduction of an Examiner Frame. The examiner frame is a script for the examiner's role in the conversation with the candidate and it guides the management of the test as it progresses through each of the three parts. The wording in the frame is carefully controlled in Parts 1 and 2 to ensure that all candidates receive similar input. In Part 3, the two-way discussion, the frame is looser and the examiner can accommodate their language to the level of the candidate by fashioning appropriate questions from graded prompts. The frame also provides support for lower level candidates while still allowing higher level candidates the opportunity to demonstrate their proficiency. Once again this is an important feature in maintaining the prescribed timing and format of the test, and in achieving standardisation worldwide. The examiner frame for IELTS was developed on the basis of what already happens during the current IELTS Speaking Test and was also informed by UCLES' extensive experience of speaking test development for its other examinations.

In September 1998 selected sample tasks for all three test parts were trialled on a small scale with both native speakers and potential IELTS candidates to check timings, clarity of instructions, feasibility, etc; some performances were videoed for the purpose of analysis and discussion by the Working Party. The revised content specification and the results of initial trialling were then discussed at a series of IELTS consultative meetings held in Australia and the UK in October and November 1998.

Following extensive consultation with the IELTS partners, administrators, Chief Examiners, Senior Examiners, and invited experts working in the field of oral proficiency assessment, the content specification and sample tasks were further refined and revised to produce a set of materials and procedures (including an examiner frame) for more extensive trialling. These materials and procedures were used with 15 IELTS candidates by 3 IELTS Senior Examiners in Australia during January and February 1999. The Senior Examiners provided a feedback report on the trialling exercise and all the trialled performances were audio-recorded for subsequent review and analysis.

Phase 2 : Development (Jan-Sept 1999)

During March/April 1999 three Senior IELTS Examiners were commissioned to develop additional sets of speaking test materials according to the new format. These tasks were extensively edited by the Working Party in August/September 1999 and feedback from the item writers was also reviewed. Further trialling of the revised format took place and these tasks have since been used in the production of sample materials, training/certification materials and live test versions.

Phase 3 : Validation (Oct 1999-Sept 2000)

Phase 3 involved an experimental study to investigate more closely the functioning of the redesigned test format, primarily the nature of the candidate output but also the handling of the Examiner Frame. A set of 39 trial tests were administered and audio-recorded by 4 Senior IELTS Examiners in the UK and Australia; examiners used the same set of materials for each test. A dataset of 20 recordings were selected for transcription and analysis: this subset included 13 female and 7 male subjects, scoring between Band 3 and Band 8, and represented 11 different L1s. Results from this study (Lazaraton, 2000) suggested that the revised test format is capable of routinely eliciting a broad range of speaking functions. Table 2 lists the different speaking functions which were identified as emerging regularly across the 20 performances.

Table 2: Speaking functions easily identified and regularly occurring in the data

Providing personal information	Suggesting	Summarising
Providing non-personal information	Justifying opinions	Conversation repair
Expressing opinions	Speculating	Narrating and paraphrasing
Explaining	Expressing a preference	Analysing
	Comparing and contrasting	Qualifying

It is worth noting that additional functions to those listed above may occur in the course of the speaking test but that they cannot necessarily be forced or predicted by the test structure, e.g. asking for information/opinions, agreeing/disagreeing.

In the light of the findings from this study, further minor adjustments were made to the specifications, to the examiner frame and to the planned content of the examiner training materials.

Phase 4 : Implementation (Oct 2000 – June 2001)

Phase 4 of the project focused on the production of material for future live test versions, for the new specimen materials and for the examiner training/certification program. This phase also included the retraining program for all IELTS examiners worldwide in readiness for July 2001; a report on this activity will be included in the next issue of *Research Notes*.

Phase 5 : Operational (from July 2001)

The revised IELTS Speaking Test became operational in July 2001 at all IELTS centres throughout the world. Following introduction of the revised test format, candidate score and test performance data will continue to be systematically gathered in order to monitor the functioning of the test. The fact that IELTS speaking tests are routinely recorded onto cassette for checking and monitoring purposes also makes it easier to undertake studies of the spoken production of both examiner and candidate, through the use of observation checklists (see *Research Notes 2* and *3*) or through analysis of more detailed transcripts of performance as reported here.

In conclusion, the test format and tasks in the revised IELTS Speaking test have been designed with a bias for best for *candidates*, who will have more opportunity to speak at length and display their ability in English more fully than was previously possible; and for *examiners*, who now have a more user-friendly and standardised brief. Throughout the revision project trialling feedback from both candidates and examiners has been generally positive. Candidates have commented favourably on the broader range of topics, the opportunity to speak at length in Part 2, and the clear, more formal structure of the test. Examiners comment on the increased quantity and range of language now produced by candidates, especially during the long turn which allows the examiner to sit back and listen; they also appreciate the extent to which the Examiner Frame provides them with content support and so allows them to focus their attention on assessing rather than thinking what to say next. The ongoing IELTS validation program will include a survey in 2002 of candidate and examiner reactions to the revised test once it has become established.

References and further reading

Brown, A and Hill, K (1998): Interviewer style and candidate performance in the IELTS oral interview, in *IELTS Research Reports* – Volume 1, Elicos Association/IELTS Australia Pty Ltd

Garman, M (1990): *Psycholinguistics*, Cambridge University Press.

Ingram, D E and Wylie, E (1993): Assessing speaking proficiency in the International English Language Testing System, in D Douglas and C Chappelle (eds) *A New Decade of Language Testing*, TESOL, Inc

Lazaraton, A (2000): An analysis of the relationship between task features and candidate output for the revised IELTS speaking test. UCLES EFL Internal report.

Lazaraton, A (2001): *A qualitative approach to the validation of oral language tests*, Studies in Language Testing 14, UCLES/CUP

Levelt, W (1989): *Speaking: from Intention to Articulation*, Cambridge, Mass.: MIT Press.

Merrylees, B and McDowell, C (1999): An investigation of speaking test reliability with particular reference to examiner attitude to the speaking test format and candidate/examiner discourse produced, *IELTS Research Reports* – Volume 2, IELTS Australia Pty Ltd

Skehan, P and Foster, P (1997): The influence of planning and post-task activities on accuracy and complexity in task-based learning, *Language Teaching Research*, Volume 1/3, pp 185-211

Tonkyn, A and Wilson, J (forthcoming): *Revising the IELTS Speaking Test*, in the proceedings of the BALEAP 2001 Annual Conference, Strathclyde

Wigglesworth, G (1997): An investigation of planning time and proficiency level on oral test discourse, *Language Testing*, Volume 14/1, pp 101-122.

The ALTE Can Do Project and the role of measurement in constructing a proficiency framework

Neil Jones, Senior Research and Validation Co-ordinator

Through the Framework Project ALTE members have classified their examinations within a common system of levels, with the aim of promoting the transnational recognition of certification in Europe. Part of this effort, the Can Do Project (see *Research Notes 2* for an introduction), aims at providing a comprehensive description of what language users can typically do with the language at each level, in the various language skills and in a range of contexts. The Can Do Project has a dual purpose: to help end users to understand the meaning of exam certificates at particular levels, and to contribute to the development of the Framework itself by providing a cross-language frame of reference.

This article provides an update on progress, and also attempts to draw some conclusions from a phase of the work which is nearing completion: the calibration of the individual Can Do statements on the basis of empirical data from self-report questionnaires.

Calibration means establishing the precise difficulty of each statement, in relation to a single scale, so that the Can Do statements become a yardstick against which any learner or any language exam can be measured.

Structure of the Can Do scales

The Can Do scales consist currently of about 400 statements, organised into three general areas: *Social and Tourist, Work*, and *Study*. Each area is sub-divided into a number of more particular concerns, e.g. the Social and Tourist area has sections on *Shopping, Eating out, Accommodation* etc. Each of these includes up to three scales, for the skills of *Listening/Speaking, Reading* and *Writing*.

Each such scale includes statements covering a range of levels. Some scales cover only a part of the proficiency range, as of course there are many situations of use which require only basic proficiency to deal with successfully.

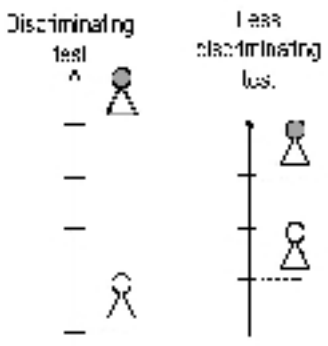
Measurement and judgement

The empirical work to make the Can Do scales into an instrument of measurement followed on from earlier work in which the Can Do

statements were constructed and assigned to levels through a process of qualitative analysis, or judgement. The aim of the empirical study – to validate, improve and add precision to the scales – has been achieved, and yet an important conclusion of this work is that measurement and judgement are complementary and equally important aspects of constructing a proficiency scale.

Life provides enough illustrations of the shortcomings of judgement. The shortcomings of measurement are less apparent, and so they will be presented in what follows.

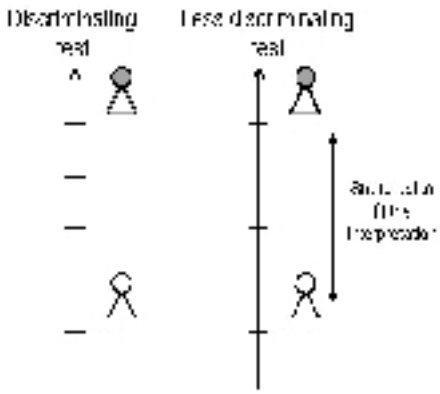
Let us begin by asking: how many levels of language proficiency are there? The ALTE Framework has five, or six if we include the embryo Breakthrough level. In this it agrees with the Council of Europe Common Framework, at least at one level of sub-division. Six is a reasonable number: large enough for putting learners into groups of practically comparable ability, and small enough to make distinctions of practical significance. But from a strictly measurement point of view, as many levels exist as a given measurement instrument is able to distinguish. A very short placement test might reliably distinguish just three levels, whereas a very long and time-consuming assessment might distinguish ten or more. In measurement terms, one unit on a proficiency scale means one reliably distinguishable shade of ability. Figure 1 shows two tests of varying discrimination: On one the two learners are separated by four units, on the other by only two.



Two tests of varying discrimination

Figure 1

We might interpret this to mean that the pair on the right are closer in ability. But suppose we know that the pairs on the left and right are in fact the same people, who have taken both tests. The natural interpretation which we will probably wish to impose is that each learner has a single ability level, as shown in Figure 2. To fit the shorter scale to this interpretative framework we have to stretch it out.



Two tests forced to fit an interpretative framework

Figure 2

Note that in this situation we have a basis for identifying the different discrimination of the tests, and bringing them into agreement with each other. But where our responses come from different groups of people (as with the Can Do data), it becomes much more difficult to distinguish between substantive differences in ability and differences in the precision of measurement of scales.

Why do tests (or Can Do scales) vary in their capacity to measure? Other things being equal, a longer test will always discriminate better. But other things are often not equal. It is a fact of assessment life that some things are more measurable than others. Thus scales that measure some aspects of language proficiency turn out shorter than others – the two figures above illustrate, for example, a situation actually observed in an oral interview procedure using separate scales for measuring grammatical accuracy and pronunciation. The pronunciation scale tends to be less discriminating. In this case it is human raters who are able to distinguish one aspect more finely than the other. In objective tests, it is the tendency of learners to

respond uniformly to items – that is, to agree on what is difficult or easy, relative to their level - that makes for precise measurement.

The Can Do statements as a measurement instrument

It is central to the idea of using the Can Do statements to define a framework of levels that people will agree on what is difficult or easy. If users of a foreign language, irrespective of what that language is, or of their own language, or their educational or professional background, agree on how they rank tasks by difficulty, then there is a basis for using such statements to describe levels in a way that will support precise measurement, and generalise well across a variety of situations of use.

A proficiency framework defined briefly and vaguely will generalise to every situation but be of no practical use. Conversely, a framework defined in great detail is unlikely to generalise across all learners or situations of use. Clearly there are limits to generalizability, and the empirical work on the Can Do project has been useful in enabling us to explore these limits. In the end our aim remains to construct a useful, practical descriptive framework.

The Can Do data collected so far have shown a number of effects where groups of people disagree to a greater or lesser extent on what they find easy or difficult. The next sections review some of these effects, looking first at issues associated with the text of the statements themselves, and then at issues associated with particular groups of respondents.

Textual effects

The Can Do statements exist in 13 languages, and it is not surprising that some *translation* effects were found at an early stage. Statements which were unexpectedly hard or easy when presented in a particular language were studied, and in some cases this could be linked to the translation. Another predictable effect concerned the *orientation* of statements. Negatively worded statements (Cannot Do) performed badly for higher-level respondents, who found it unnatural to endorse low-level statements negatively worded. Such statements were re-worded positively or removed.

Other effects were unexpected but interpretable. For example, *detailed exemplification* of a task tended to make it more difficult than predicted.

All these effects could be corrected to the extent that they were problematic. Somewhat more difficult to deal with were systematic differences in discrimination. This issue arose during a study to equate the Can Do scales to the Council of Europe Common Framework. Several scales from the Framework document (Council of Europe 1996) were included in versions of the Can Do questionnaires and response data collected. Although there was close agreement between self-ratings on the two types of scale, the CE statements were found to define a significantly longer scale. This was because the CE scales consisted of six detailed composite statements, each epitomizing one level, whereas the Can Do scales in their deconstructed form consisted of a larger number of short atomic statements. A satisfactory equating of the two scales required a qualitative study – that is, the exercise of judgement.

Person effects

Effects found for groups of respondents are particularly interesting, as they indicate the limits to which Can Do statements generalize across learners and situations of use.

Demographic effects studied included *age, background* and *profession.*

Respondents included a limited number between the ages of 13 and 18. This age group tended to respond in ways which were inconsistent with the responses of older people. This is not surprising, as the Can Dos chiefly concern ability to operate in an adult world, and refer to tasks which children of school age would have had no experience of.

It was found that a person's occupation or professional status might affect their ability to use a foreign language in particular situations. Thus for example, employers found it significantly easier than employees at middle or junior level to deal with situations likely to arise in a hotel, restaurant, a bank, or while travelling. This is hardly surprising. More unexpected was that employers found it significantly easier to understand a photo-copier or fax machine. What this appears to reflect is a different understanding of what the Can Do statement actually means (the employer probably has never tried to understand the instructions to fix a paper jam or change the toner cartridge).

Grouping respondents by *target language*, an interesting contrast was found between what learners of English and French find relatively hard or easy.

Whatever their overall level, learners of French seem likely to be relatively more confident of their receptive language skills (e.g. *CAN understand the general outline of a guided tour…*). Learners of English on the other hand are relatively more confident of their active communicative skills (e.g. *CAN participate in casual conversation over the phone with a known person on a variety of topics*).

It is interesting to consider what this might indicate about people's reasons for studying foreign languages or perhaps approaches to teaching different languages.

A noticeable effect concerned *ability* level. In self-report data respondents at lower proficiency levels tend systematically to over-rate their ability. That is, they have a different understanding of "can do".

Constructing a language proficiency framework: quantitative and qualitative aspects

The empirical, statistically-based approach to validating the Can Do statements has been useful in calibrating the individual statements and constructing the individual scales. It has also been useful precisely because it identifies issues with how people understand and use assessment scales. We have found evidence of a range of group effects such as age, proficiency level, or area of language use, that may affect understanding of a scale and of the meaning of Can Do level descriptors. Thus there are scales where there is good agreement as to what is hard or easy, and scales where agreement is less. Consequently, (as explained above) some scales measure more precisely than others.

Some exercise of judgement becomes necessary in order to impose a single frame of reference on the different scales. The approach followed was based firstly on a close analysis of the text of each statement, in order to identify tasks which are very similar in different areas of use (Social and Tourist, Work and Study). These were posited to be of similar difficulty. Secondly, reference was made to the ALTE levels originally assigned to statements (restricting attention to those statements which had not been edited during the textual revision). The correlation between original and empirically found level was high, and so it could be assumed that overall these assigned levels could be used to anchor the scales to each other. From these two sets of observations, a separate linear transformation was found (that is, a formula for "stretching" the scale) for each language skill within each area of use. After applying these the textual analysis was repeated. Some apparent anomalies remained in several of the scales from the Study area of use, and these were individually rescaled to bring them into line.

A subsequent step has been to select from the individual statements and construct Can Do scales consisting of composite level statements. Statements were selected both for their content and for their statistical properties. As far as possible statements about which respondents disagreed were excluded. These composite statements are used in the computer-based Can Do self-assessment tool, and will also be exploited in validation activities currently being planned.

Reference:

Council of Europe (1996): *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference*. CC-LANG (95) 5 rev IV, Strasbourg, Council of Europe.
See also Jones (2001): Appendix D – ALTE Can Do Statements, in Council of Europe (2001): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Towards a common scale to describe L2 writing performance

Roger Hawkey, Consultant, UCLES EFL

Direct tests of language performance in writing, long regarded as an important contributor to profiles of learner communicative competence (see Hawkey 1982, Weir 1993, North 2000) are an increasing focus of language-testing research, as technological advances facilitate corpus-based studies, using machine-readable text and concordancing analysis. This paper reports on a corpus-based study exploring answers to the following questions: *What are the distinguishing features in the writing performance of EFL/ESL learners or users taking the Cambridge English examinations? How can these be incorporated into a single scale of bands, that is, a common scale, describing different levels of L2 writing proficiency?*

The envisaged common scale for writing entails a set of performance descriptors expressed in terms of criterial features and applicable to all the levels of writing proficiency of English language learners and users. As a *common* scale, it would be able to identify, for example, the comparative levels of proficiency of candidates for all five examinations of the Cambridge main suite, namely the Key English Test (KET), the Preliminary English Test (PET), the First Certificate in English (FCE), the Certificate in Advanced English (CAE) and the Certificate of Proficiency in English (CPE), described in the relevant UCLES *Handbooks* as "a series of examinations with similar characteristics, spanning five levels". The five levels, represented by candidates able to pass each of the main suite exams, coincide as follows with the framework of levels of proficiency of the Association of Language Testers in Europe (ALTE).

Table 1: Cambridge Main Suite and ALTE Levels

ALTE Levels	UCLES Main Suite Exams
5 Good User	CPE
4 Competent User	CAE
3 Independent User	FCE
2 Threshold User	PET
1 Waystage User	KET

The ALTE level descriptions are essentially *user*-oriented, that is predicting social and workplace functions that those who have achieved a particular level of proficiency should be able to perform. Their broad communicative tone is illustrated by this short excerpt from the quite detailed (c. 120 word) ALTE Level 3 can-do specification:

> *In social and travel contexts, users at this level can write short notes and messages and simple personal letters of a narrative or descriptive type, such as 'thank-you' letters and post cards. In the workplace , they can write a short note of request and record a routine order. They can make notes during a meeting for their own purposes , and write a straightforward routine letter, although this will need to be checked by a colleague …*

The research in progress described here is a part of the second phase of the Common Scale for Writing (CSW) Project, initiated by UCLES EFL Validation Unit in 1994, and seeking a set of band descriptors for use by assessors of writing proficiency, as well as by learners or employers. A validated common scale would, for example, enable testers to compare directly the different written performance of an FCE and a CPE candidate, or the different typical ranges of performance across all five exams, each exam with its own benchmark level (■ in Figure 1 conceptualising the relationship between a common scale and the levels of the main suite examinations).

A common scale across examination levels and ranges

Figure 1

Research into a common scale for writing is part of UCLES' Framework Project to establish five proficiency levels matching KET, PET, FCE, CAE and CPE. The CSW Project runs parallel with UCLES research into a common scale for speaking. The new phase of the CSW project also relates to several of the key areas of UCLES EFL research identified by Mike Milanovic in 1996 in his *Studies in Language Testing* series editor's note and reiterated in the editorial article in the first issue of *Research Notes* (March 2000). These are: the direct assessment of spoken and written skills; the rationalisation of data capture; and the triangulation of test content, candidate background and test performance.

In Phase 1 of the CSW Project, existing writing assessment scales were used to derive a draft set of "pass-level" descriptors of the writing proficiencies of candidates from KET through to CPE (Capel, 1995). Liz Hamp-Lyons, also reporting in 1995, studied typical main suite candidate scripts from all five levels and proposed "can do", "can sometimes do", and "cannot do" statements with reference to assessor- as well as user - oriented features such as: task completion; communicative effectiveness; syntactic accuracy and range; lexical appropriacy; chunking, paragraphing and organisation; register control; and personal stance and perspective. The sample scripts used, which had been selected across main suite exams, naturally revealed different salient aspects of writing ability for different tasks. In order to control the task variable in the latest work on the development of a common scale for writing, all candidate scripts in the Phase 2 corpus are in response to the same task.

To obtain scripts for the new corpus, a task suitable for FCE, CAE and CPE candidates needed to be selected. After consideration of the report from the Principal Examiner on the December 1998 session, the question *Do you prefer listening to live music or recorded music?* was chosen, and a representative sample of 108 live test scripts was selected from the December 1998 FCE administration. Scripts on the same task from candidates from the two other levels of the main suite (CAE and CPE) were also needed for common scale research. A list of teaching and exam centres was thus compiled with the aim of identifying and selecting approximately 200 candidates, split equally between CAE and CPE and representing a range of regional backgrounds. In the event, six centres were sent pilot test papers with the instruction that the test, including the writing task already completed by the FCE candidates, should be administered within a 2-week period in April 1999. Experienced UCLES writing examiners were invited to a marking day at UCLES in May 1999 and oriented to the relevant FCE marking scheme. The 180 pilot CAE/CPE scripts were split into 18 batches of 10 and three mark sheets were produced for each batch. Each batch was marked by three different examiners, all using the FCE mark scheme previously used by the FCE markers for the same task in the live FCE test. Once the 180 pilot CAE/CPE scripts had been marked, photocopies of the 108 live FCE candidate scripts were remarked by members from the same team of writing examiners.

The research question to be answered in Phase 2 of the CSW project is: *Using a text-based analysis, what distinguishing features in performance can be identified among writers across three proficiency levels addressing a common task?* Seeking insights on this, Roger Hawkey, appointed by UCLES to co-ordinate Phase 2 of the CSW Project, first rated and sorted into performance-quality groups all 288 scripts in the corpus. This process was carried out without his prior knowledge of whether the scripts had been written by FCE-, CAE- or CPE-level candidates, or of the ratings already assigned by the UCLES raters. Hawkey's rating and sorting was done using the current FCE rating scale, and drawing on his own view as a 'reader' together with a 'naïve' understanding of how to apply the scale. As well as rating all the scripts, brief comments were written on each in terms of its distinguishing features of performance.

At the end of this first stage of the investigation, a dataset was created of all markings of the 180 pilot CAE/CPE scripts and the remarking of the 108 live FCE scripts; the agreement between the ratings assigned by the script analyst and those of the experienced UCLES raters was checked. Table 2 shows inter-rater correlations high enough not to invalidate findings based on the groupings of scripts according to the FCE assessment band scores assigned to them.

Table 2: Inter-rater correlations

		Live mark	Error count	Corpus analyst Score	Mean score of all pilot markings
Live mark	Correlation	1.000	-.557	.774	.638
	Sig. (2-tailed)	.	.000	.000	.000
	N	108	108	108	108
Error count	Correlation	-.557	1.000	-.759	-.606
	Sig. (2-tailed)	.000	.000	.	.000
	N	108	288	288	288
Corpus analyst score	Correlation	.774	-.759	1.000	.810
	Sig. (2-tailed)	.000	.000	.	.000
	N	108	288	288	288
Mean score of all pilot markings	Correlation	.638	-.606	.810	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	108	288	288	288

Table 3: Corpus and sub-corpus rater statistics

	Number of candidates	Mean	SD
Score from first pilot marking	288	15.18	2.78
Score from second pilot marking	180	15.47	2.75
Score from third pilot marking	180	15.67	2.61
Live FCE exam mark	108	14.49	3.29
Corpus analyst score	288	15.12	2.64
Mean mark from pilot markings	288	15.27	2.51
Difference between corpus analyst's score and mean pilot	288	-0.16	1.59

The ratings, including those of the script analyst – Roger Hawkey, were now used in the selection of three sets of scripts, the first set being all scripts banded at 5 by all raters on the FCE scale (n=29), the second set (n = 43) banded at 3 by all raters, and the third (n = 8 only) at band 2 by all raters. Since the scripts in these sub-corpora had attracted unanimous rating agreements across raters, the three sub-corpora could be regarded as representing *high, medium* and *low* proficiency levels, regardless of examination candidacy.

Each of the sub-corpora scripts was then submitted by the script analyst to detailed re-examination involving:

- a re-reading
- an error count using conventional teacher error categorisations
- a characterisation according to main communicative descriptors, i.e. what strikes the reader as good and not so good
- a selection of script extracts considered 'typical' of communicative characteristics of the sub-corpus, i.e. level

Early findings from this close analysis of all the scripts in the three sub-corpora suggest that the three levels of proficiency are consistently distinguished by the way in which the following features *impact on the reader*:

- 'sophisticated' language use, as manifested, for example, through advanced vocabulary, collocation, idiom, pace variation, humour
- frequency and type of linguistic inaccuracy
- clear organisational structure and effective, natural links.

A distinguishing feature along the lines of 'balance of personal experience and objective argument' could also be included in the draft band descriptors, but the script corpora need further analysis on this point.

This analysis of the three sub-corpora permits the development of can-do descriptors for each of the identified key features of different levels of written proficiency. Evidence from the three sets of scripts on the 'sophisticated language' characteristic, for example, suggests common scale can-do band descriptions such as the following:

HIGH LEVEL:
CAN WRITE WITH IMPACT AND APPROPRIATE STYLE ON NON-SPECIALIST DISCUSSION TOPICS USING EFFECTIVELY: ADVANCED VOCABULARY, COLLOCATION, WORD ORDER, IDIOM, PACE VARIATION AND/OR HUMOUR.

MID-LEVEL:
CAN WRITE ON NON-SPECIALIST DISCUSSION TOPICS BUT WRITING GENERALLY LACKS IMPACTFUL AND STYLISTICALLY APPROPRIATE USE OF VOCABULARY, COLLOCATION, WORD ORDER, IDIOM, PACE VARIATION AND/OR HUMOUR

LOW-LEVEL:
CAN WRITE ON NON-SPECIALIST DISCUSSION TOPICS BUT WRITING LACKS IMPACTFUL AND STYLISTICALLY APPROPRIATE USE OF VOCABULARY, COLLOCATION, WORD ORDER, IDIOM, PACE VARIATION AND/OR HUMOUR

The feature "frequency and type of linguistic inaccuracy" suggested above may appear somewhat out of tune with current communicative assessment criteria, where appropriacy is generally felt to take precedence over accuracy. An analysis of the scores given by the experienced raters in the Phase 2 study, however, seems to underline the importance of accuracy as a criterion and distinguishing feature in the rating of writing performance, while by no means exaggerating it. Table 4 compares the numbers of errors in the scripts of the sub-corpora (n=80) as noted by the script analyst using standard FCE annotation of 'errors relating to general control of language or specific marking points, omissions, doubtful words or phrases, wrong order', with the band scores assigned by all raters. A significant, though not exclusive, relationship between accuracy and overall score is indicated, i.e:

Table 4: Accuracy error frequencies across sub-corpora

Sub-Corpus	Average length (no of words)	Error No. range	Average accuracy errors marked	Error Frequency (per No. of words)
Band-5	210	1-11	5	1 per 42
Band-3	188	11-31	17	1 per 11.1
Band-2	154	8-39	24	1 per 6.4

There are also interesting signs in the Phase 2 research that the negative impact made by certain accuracy error types is greater than that made by others, or that students with weaker target language competence are more likely to make certain types of accuracy errors. Common scale descriptor elements to cover formal accuracy could be along the following lines:

HIGH LEVEL:
CAN WRITE (ON NON-SPECIALIST DISCUSSION TOPICS) WITHOUT BASIC ERRORS OF GRAMMAR OR VOCABULARY REDUCING IMPACT ON THE READER, EVEN IN TASKS WHERE ACCURACY OF GRAMMAR AND VOCABULARY ARE CONSIDERED IMPORTANT

MID LEVEL:
CAN WRITE ON NON-SPECIALIST DISCUSSION TOPICS BUT IMPACT ON THE READER WILL SOMETIMES BE REDUCED BECAUSE OF BASIC ERRORS OF GRAMMAR OR VOCABULARY, ESPECIALLY IN TASKS WHERE ACCURACY OF GRAMMAR AND VOCABULARY ARE CONSIDERED IMPORTANT

LOW LEVEL:
CAN WRITE ON NON-SPECIALIST DISCUSSION TOPICS BUT IMPACT ON THE READER WILL BE SIGNIFICANTLY REDUCED AND UNDERSTANDING OF THE INTENDED MESSAGE OCCASIONALLY INHIBITED BY FREQUENT BASIC ERRORS OF GRAMMAR OR VOCABULARY.

The corpus analysis also suggests that the organisation of the points in the argument in response to the invitation to the candidates to give their written views on live versus recorded music, and the way they link their points provide further systematic differences by proficiency level. Noting the clear evidence that discourse cohesion and coherence markers can be over- as well as mis-used, the following descriptors begin to reflect the performance on organisation and links in the three sub-corpora.

HIGH LEVEL:
CAN WRITE (ON NON-SPECIALIST DISCUSSION TOPICS) WITH IMPACT ENHANCED BY A CLEAR ORGANISATIONAL STRUCTURE USING EFFECTIVE AND NATURAL LINKS, IMPLICIT AS WELL AS EXPLICIT, AND REACHING LOGICAL CONCLUSIONS

MID LEVEL:
CAN WRITE ON NON-SPECIALIST DISCUSSION TOPICS BUT WEAKNESSES IN THE ORGANISATIONAL STRUCTURE AND SOME INEFFECTIVE AND/OR OVER-EXPLICIT LINKS REDUCE THE IMPACT OF ANY CONCLUSIONS

LOW LEVEL:
CAN WRITE ON NON-SPECIALIST DISCUSSION TOPICS BUT WEAKNESSES IN THE ORGANISATIONAL STRUCTURE AND MISUSED LINKING DEVICES MAY OBSCURE AND/OR WEAKEN THE ARGUMENT AND CONCLUSIONS

Such descriptor elements, once further validated, will be combined with other distinguishing features of writing proficiency and divided appropriately into the number of levels targeted. Collaboration with related corpus analytical research will provide further validation of the draft scale band descriptors. (It is worth noting that the high-level can-do descriptions have already proved valuable in informing the revision of the writing assessment descriptors for the CPE Revision Project.)

One collaborative project already under way (under the IELTS funded research program) is with Dr Chris Kennedy of the University of Birmingham, who has been carrying out, with colleagues, an analysis of an IELTS corpus (N=150). The methodology involves trans-processing

candidates' scripts into Word, including all errors, performing a 'manual analysis' to note features of interest by band level, performing statistical analyses on essay length, and then using the Concord and Wordlist tools of Wordsmith for data on word frequency, concordances, and collocates. Interesting features are already emerging from this research, some of them clearly reinforcing some of the early findings of the CSW Project, e.g.

- longer essays with broader vocabulary range at higher IELTS proficiency levels;
- more rhetorical questions, interactivity, idioms, colloquial and colourful language, metaphor, at higher levels, while almost none at lower levels;
- the use of too many explicit cohesion devices by candidates at lower writing proficiency levels.

In a new project proposal, Kennedy (2001, p1) suggests the following link with the UCLES CSW research:

> 'The aims are to transfer the existing corpus of written answers developed by Hawkey *et al* (FCE/CAE/CPE levels) to machine readable form so that the linguistic nature of the levels of performance may be analyzed. The outcomes will be the creation of a performance database and a linguistic specification of the three levels selected, that can be compared with the work already done by Hawkey et al.'

This collaboration with the Birmingham team should strengthen both research projects, ensuring a convincing combination of qualitative and quantitative methodologies to analyse L2 writing performance.

The initial work of Phase 2 of the CSW Project was presented at a research-in-progress session at the LTRC 20001 Conference in St Louis, USA, in February 2001 and will be submitted for journal publication in the summer of 2001.

Acknowledgements

We would like to acknowledge the involvement of Nick Saville, Janet Bojan, Annette Capel and Liz Hamp-Lyons in Phase 1 of the CSW Project. Thanks to Roger Hawkey, Nick Saville, Chris Banks and Beth Weighill for their work to date on Phase 2.

References

Capel, A (1995): *Common Scale for Writing Project*: UCLES EFL Internal report.

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

Hawkey, R (1982): An investigation of inter-relationships between cognitive/affective and social factors and language learning. Unpublished PhD thesis: Department of English for Speakers of Other Languages, Institute of Education, London University.

Kennedy, C and Dudley-Evans, T (forthcoming). *Investigation of linguistic output of (IELTS) Academic Writing Task 2*: Centre for English Language Studies, and English for International Students Unit, University of Birmingham, England.

Kennedy, C (2001): *Investigation of linguistic output of three levels of test performance*, IELTS research proposal.

Milanovic, M, Saville, N and Shen, S (1992): *Studies on direct assessment of writing and speaking*: UCLES EFL internal report.

North, B (2000) Linking language assessments: an example in a low-stakes context. *System* 28: 555-577.

Saville, N and Capel, A (1996): *Common Scale for Writing, Interim Project Report*: UCLES EFL Internal report.

Stewart, M. and C, Grobe (1979): Syntactic maturity, mechanics of writing and teachers' quality ratings, *Research in the Teaching of English* 13: 207-15.

Taylor, L (2000): EFL research at UCLES, in *Research Notes* 1, p 2.

Weir, C (1993): *Understanding and Developing Language Tests*. Hemel Hempstead: Prentice Hall.

CB BULATS: Examining the reliability of a computer based test using test-retest method

Ardeshir Geranpayeh, Research and Validation Officer, UCLES EFL

Introduction

The stability of test results over time has been one of the concerns of test designers. One way of demonstrating that stability is by means of test-retest, where a group of candidates sit for the same test twice over a period of time. The Pearson correlation between the scores on the two sittings is called the stability coefficient and is indicative of the reliability of the test. A coefficient of 0.80 or more would generally indicate that the data are reliable enough for practical purposes. Although the stability coefficient is the most appropriate way to show the stability of test results over time, it is not very often reported in language testing literature. This is because it is very difficult to persuade a group of test takers to sit for the same test twice and expect them to take the exam with the same degree of attention on both occasions. This short paper examines the reliability of a computer based test using the test-retest method. The current study follows up the work reported on BULATS by Neil Jones in *Research Notes 3* (November 2000), where a computer-based version of BULATS was compared with the paper-and-pencil version. Jones' study demonstrated that there was a linear relationship between the CB and P&P scores, supporting the view that it should be practical to develop the two formats for use interchangeably. The reliabilities reported for the P&P format and that of the CB were 0.93 and 0.94 respectively and the correlation between the scores on the two tests was 0.86 when six outlying cases were removed. Based on the square of alpha reliability, the study predicted that we would get a correlation of 0.88 between the scores on two sittings of the CB format. The accuracy of such a prediction will be examined in this report by estimating the reliability of a CB test using both the stability coefficient and a Rasch reliability estimate (an internal consistency measure, analogous to Cronbach's Alpha).

The CB BULATS test-retest project

CB BULATS is currently under revision and a new version of the test will be released shortly. The new version, while maintaining the adaptive mode, includes new item types and is relatively longer. As part of the validation exercise, the new version of the test was piloted in Cambridge earlier this year. The main objective of the project was to examine the stability of the new CB BULATS test scores over time using the test retest method. Other issues to be investigated were:

1 The effect of an adaptive mode of administration on test reliability and discrimination, and

2 The effect of test taker features such as L1, gender, age, and familiarity with computers on test scores.

Administration

87 EFL test takers studying at various language schools in Cambridge volunteered to take the new version of CB BULATS twice on the same day with a short break between the two administrations. They were also given a questionnaire to complete. 85 test takers completed the questionnaire. Table 1 demonstrates how the test takers varied with respect to their L1, gender and age.

Table 1: Test takers grouping by L1, Gender & Age

Grouped by L1 Language

First Language	Frequency	Percent	Cumulative %
Arabic	1	1.18	1.18
Chinese	1	1.18	2.35
Faeroes	1	1.18	3.53
French	3	3.53	7.06
German	8	9.41	16.47
Italian	3	3.53	20.00
Japanese	5	5.88	25.88
Missing	4	4.71	30.59
Portuguese	8	9.41	40.00
Russian	1	1.18	41.18
Slovak	1	1.18	42.35
Spanish	44	51.76	94.12
Turkish	5	5.88	100
Total	85	100	

Grouped by Gender

GENDER	Frequency	Percent	Cumulative %
Female	55	64.71	64.71
Male	30	35.29	100
Total	85	100	

Grouped by Age

Age Group	Frequency	Percent	Cumulative %
10-16	9	10.59	10.59
17-20	43	50.59	61.18
21-25	16	18.82	80.00
26-34	14	16.47	96.47
35+	3	3.53	100
Total	85	100	

The candidates' test scores on the two sittings and their responses to the questionnaire were entered into a database for further analysis. For ease of reference, the first administration of the test will be called Test 1 and the second (retest) referred to as Test 2. Reference to results reported in Jones'

study will be referred to as Test 3. Test 1 and Test 2 are the new version of CB BULATS, while Test 3 is the current version of the test.

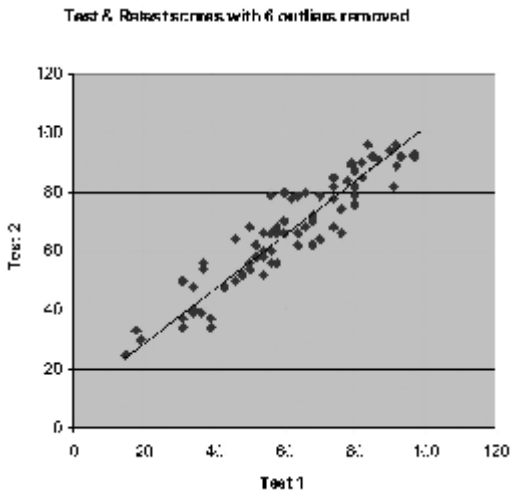
Findings

It is important to mention that test scores on CB BULATS do not refer to raw scores. They are actually ability estimates derived from a latent trait (Rasch) analysis, converted into BULATS scores by means of a scaling procedure. The items in the test and retest were taken from the same item bank with calibrated item difficulties (see Jones, *Research Notes 3* on item banking). The following terminology will be used with reference to the scores: Test Score refers to BULATS test score (0-100), Band Score refers to BULATS band scale (1-5), and Ability level refers to candidate ability as estimated by Rasch model (Logit).

Reliability

The average reliability (Rasch) for each version of the test was estimated as 0.94, and 0.93 for Test 1 and Test 2, respectively. Using the square of this reliability to model the correlation between two sittings of the test, the estimated reliability was 0.87. This figure is very close to the prediction that Jones estimated for the CB BULATS test retest coefficient in his study (0.88). The test scores from Test 1 and Test 2 were correlated to examine how accurate these predictions were. The correlation between the two test scores was 0.89 before any outliers were removed and 0.93 when six outliers were removed. The stability coefficient between Test 1 and Test 2, even before removing the outlying cases, is higher than the value that the square of the alpha reliability predicts. This allows us to be relatively confident about the stability of the new CB BULATS test scores over time.

Figure 1 shows a scatterplot of test-retest scores (with six outliers removed, i.e. replicating the approach used in the previous study). Sitting for a test twice on the same day under experimental conditions will produce variations in performance; however, the high correlation (0.93) achieved between the scores of the candidates on test-retest shows that any such variations were minimal.



CB BULATS ability scores compared in two sittings

Figure 1

Figure 1 also indicates that there is good agreement in overall level between the scores obtained on the two sittings. The spread of test scores along the identity line shows that the tests are discriminating relatively well; a similar finding was reported by Jones for Test 3. It appears that CB format, in general, can produce more discriminating results. This is due to the adaptive mode of the test, which selects the most appropriate items for each candidate according to their estimated level, providing more information per item and minimising the effect of guessing.

Table 2 : Mean and SD of band scores

	Band Scores		
	Test 1	**Test 2**	**Test 3**
Mean	2.78	3.06	2.80
SD	1.32	1.18	1.24

Table 2 reports the mean and SD of band scores of the candidates for the current and new version of CB BULATS. The mean band scores on Test 1 and Test 3 and their variability in scores (SD) are so close that it allows us to conclude that the two populations were similar in terms of their ability. The slight change of band scores in Test 2 is due to the better performance of the test takers on their second attempt. To determine whether the differences in candidates' test scores / bands in Test 1 and Test 2 were significant, t-tests were applied.

Table 3 : Results of tests of significance

Variables compared		t	df	Sig. (2-tailed)
Pair 1	BandScore1- BandScore2	-3.396	86	0.001
Pair 2	TestScore1 - TestScore2 -	4.375	86	0.000

Table 3 illustrates that the candidates scored significantly higher in retest. This improvement in language ability was greatest for lower-level candidates, hence the lower SD of scores observed for Test 2. We will be discussing this in *Final Remarks*.

Table 4 compares Rasch reliability estimates for the three CB BULATS tests. The current version (Test 3) and the new version (Test 1 & Test 2), despite their differences in format and length, seem to be comparable with respect to their mean standard error of measurement, separability and reliability (Rasch) estimates. The slight decrease in the reliability of Test 2 is due to the better performance of test takers on retest, which resulted in lower variability in scores on their second attempt.

Table 4 : Test Reliabilities (Rasch)

	Test 1	Test 2	Test 3
Ability SD	1.29	1.12	1.32
Mean SEM	0.33	0.31	0.33
Separability	3.90	3.61	3.99
Reliability	0.94	0.93	0.94

Test 1 & Test 2 (Test & Retest) = New version of CB BULATS (this study)
Test 3 = Current version of CB BULATS (reported in Jones' study)
Ability SD= Standard Deviation of candidate's ability
SEM =Standard Error of Measurement

The effect of test taker features on test scores

There are various ways of examining the influence of test taker features such as gender on test results of which Analysis of Covariance (ANCOVA) is one. ANCOVA is a means of reducing systematic bias, as well as within-groups error in the analysis. The aim is to determine whether the independent variable – gender, age, etc. – is indeed having an effect on the dependent variable, i.e. Test 1 scores; we do this by statistically controlling the influence of an extraneous variable such as Test 2 scores (covariate) on the dependent variable. In other words, we attempt to reduce the error variance caused by individual differences.

To examine the effect of test taker features on test scores, a number of *One - Way Analysis of Covariance (ANCOVA)* were conducted on test scores with respect to the information collected through the questionnaire. In each ANCOVA, test score on Test 1 was the dependent variable, test score on Test 2 was the covariate and the feature under investigation was the independent variable. Features investigated were as follows: L1, gender, age, familiarity with computers, frequency of computer use, preference in using CBT and P&P, and suffering from eye strain during the test (Test 1 & Test 2). None of the analyses conducted indicated that there was a main effect ($p>.05$) for the features examined. Thus we can say that test taker features examined in this study seem to have no influence on test scores in CB BULATS. A similar finding was also reported in Jones' study.

Final remarks

This research project followed up the work in Jones' earlier study where, amongst other findings, a linear relationship was reported between the scores of CB and P&P versions of BULATS. The main objective of the present study was to examine the stability of CB BULATS test scores over time and across versions.

We have demonstrated that CB BULATS test scores remain highly stable across versions and over time with a reliability estimate of 0.94 and a stability coefficient of 0.93. We have also shown that familiarity with computers does not seem to advantage / disadvantage CB BULATS candidates. The finding that we have overall higher test-retest agreement for CB-CB (0.93) than for CB-P&P (0.86), however, may indicate that the mode of administration has an effect. This will be addressed in future issues of *Research Notes*.

Finally, we have observed that the candidates scored significantly higher in their second attempt, which might indicate practice effect. Observation of individual cases shows that the variation is greatest in the scores of lower-level candidates. It could be that some of the candidates did not know how or when to key their responses; having done the test once, they had a better sense of what was expected of them. This study did not aim at examining CB practice effect, therefore further speculation does not seem to be warranted at this stage. The practice effect of a CB test can be examined in future research projects.

[i] The Business Language Testing Service (BULATS) is a language assessment service specifically for the use of companies and organisations. The service is designed to test the language of employees who need to use a foreign language in their work, and for students and employees on language courses or on professional/business courses where foreign language ability is an important element of the course.

[ii] See Jones' article in *Research Notes 3* (November 2000), pp. 10-13, for more detailed discussion of computer adaptive testing.

Studies in Language Testing

Volume 13 in the *Studies in Language Testing* Series addresses the issue of spoken language assessment looking in particular at the equivalence of direct and semi-direct oral interviews. Kieran O'Loughlin's work is based on the development and validation of the spoken language component of the **access**: test designed in the early 1990s for migrants to Australia. It is an important language testing project in the Australian context and was funded by the Commonwealth Department of Immigration and Ethnic Affairs. While the project as a whole brought together experts from a number of Australian universities, the oral test was developed by a team at the University of Melbourne. This volume is of particular significance and interest to the language testing community because it takes a multi-faceted view of the investigation of test comparability. While much research of this sort has tended to look only at quantitative data, largely correlational analyses, O'Loughlin taps into a range of different types of evidence and attempts to explore the process of construct validation in oral assessment to a depth that is rarely found.

The assessment of spoken language ability is a topic of enduring importance in the work of the University of Cambridge Local Examinations Syndicate (UCLES) given that UCLES assesses the spoken language ability of about 800,000 candidates around the world every year. The issue of semi-direct versus direct assessment of speaking has continued to be a topic of interest at UCLES and we have found that O'Loughlin's work makes a valuable contribution to our understanding. His work closely reflects our own interests particularly in the area of the qualitative analysis of oral interview interaction.

The importance of oral assessment and the need to better understand the complex issues and interactions that underlie performance in this particular context have long been a topic of debate at UCLES. As early as 1945, Jack Roach, an Assistant Secretary at UCLES at the time, was writing on the topic in his internal report entitled 'Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students' Indeed, in his book Measured Words (1995), Bernard Spolsky considers Roach's work to be 'probably still one of the best treatments in print of the way that non-psychometric examiners attempted to ensure fairness in subjective traditional examinations'. Roach's work is addressed in more detail by Cyril Weir in a volume currently being prepared for this series that focuses on the revision of the Certificate of Proficiency in English (CPE).

Over the last ten years or so, a considerable amount of work has taken place at UCLES in order to gain a better understanding of oral interview interactions, processes and rating scales both in quantitative and qualitative studies. Working internally or with colleagues at universities in the UK, USA and Australia, numerous studies have been carried out. Amongst other things, projects have looked at:

- discourse variation in oral interviews;
- rating scale validation;
- interlocutor frames and how their use by examiners can be described and monitored;
- interlocutor language behaviour;
- a methodology to allow test designers to evaluate oral assessment procedures and tasks in real time;
- comparisons of one-to-one and paired oral assessment formats;
- test takers' language output;
- the development and validation of assessment criteria.

In 1998 UCLES EFL established, within its Validation Department, a dedicated unit to drive research in the area of performance testing, which essentially covers the assessment of speaking and writing.

It should also be noted that the next volume in this series, *A qualitative approach to the validation of oral language tests* by Anne Lazaraton, also makes a valuable contribution to the assessment of spoken language ability. Both O'Loughlin's and Lazaraton's volumes underline UCLES' commitment to furthering understanding of the dimensions of spoken language assessment.

Studies in Language Testing - Forthcoming volumes 2002

A new feature of the SILTS series in 2002 will be the publication of three volumes documenting major revision projects which have been conducted by UCLES EFL in recent years. Each volume reports on a separate project and in each case a guest editor has been invited to co-ordinate the writing up of the work in association with the internal UCLES staff involved in the projects.

The three volumes are as follows:

Volume 15 Innovation and Continuity: Revising the Cambridge Proficiency Examination – Edited by Cyril Weir

Volume 16 The Development of CELS - a modular approach to testing English language skills – Edited by Roger Hawkey

Volume 17 Issues in Testing Business English: The Revision of the Cambridge Business English Certificates – Edited by Chris Kennedy and Barry O'Sullivan.

Volume 15 – Innovation and Continuity: Revising the Cambridge Proficiency Examination

The title of this volume – innovation and continuity – has been chosen to reflect the structure of the book. The revised CPE to be introduced in December 2002 demonstrates just how much language teaching and testing has changed during the last century. Nevertheless it is important to recognise that the innovations which have been introduced by UCLES are grounded in traditions within UCLES and an approach to assessment which can be traced back to the early days of the last century.

In Chapter 1 the editor traces the history of CPE from its introduction in 1913 up to the present day and the current revision project which was initiated in 1992. Chapter 2 describes in some depth UCLES EFL's approach to test development and revision projects (cf *Research Notes 4* – Test Development and Revision) and the remaining chapters describe, paper by paper, the actual changes which have been made. There is also an extensive range of appendices with specimen papers of the examination – both the revised version and examples from the past.

CPE remains a high-level examination of English, suitable for candidates of all nationalities, from a range of backgrounds. Among the features introduced in the revised exam are:

– the inclusion of texts and extracts from a wide range of sources so that candidates will read and listen to a variety of language registers and styles
– the use of both long and short texts in the Reading, Listening and Use of English papers
– writing tasks based on real life activities which will be more meaningful to the candidates
– a paired Speaking test which encourages candidates to communicate using a variety of language functions and which enhances test reliability

While changes have been made on every paper, some characteristics of CPE will be familiar.

– The revised CPE will measure at the same level of general language ability as the current CPE and to the same standards (i.e. it will be at Cambridge/ALTE level 5 representing a very high level of language ability suitable for study purposes).
– The content of CPE will continue to be particularly appropriate for candidates with a broad language learning or study background.

UCLES published the revised specifications in Summer 2000 and the booklet contains sample papers, a sample listening paper on CD, answer keys and specimen answers.

Volume 16 – The Development of CELS - a modular approach to testing English language skills

The notion of partial competence in another language was discussed briefly in *Research Notes 3* (November 2000). It was argued that this is now an important consideration in language learning around the world where learners may, for example, acquire comprehension skills (passive knowledge in listening and reading) without productive ability.

The Certificates in English Language Skills (CELS) form a modular system of examinations which allows for English language competence in reading, writing, listening and speaking to be assessed separately. Candidates for these examinations will have the flexibility to choose to do one skill at a particular level or to build up a profile of skills at one or more levels.

This volume records how CELS was developed from the Certificates in Communicative Skills in English (CCSE) and the Oxford EFL Reading and Writing Tests. The editor traces the early developments back to the mid-1970s and describes how the early work of those involved in the *communicative testing* movement has left an important legacy which is now reflected in the revised CELS exams.

The full specifications of CELS and sample papers are now available from UCLES EFL and the first session of the revised examinations will be in May/June 2002.

Volume 17 – Issues in Testing Business English: The Revision of the Cambridge Business English Certificates

UCLES offers two complementary examination systems for work-related contexts.

a) *The Business Language Testing Service* (BULATS) is non-certificated, and offers employers a quick, reliable and flexible method of assessing employees' language skills.

b) *The Business English Certificates* (BEC) are certificated examinations at three levels which can be taken on six fixed dates per year at approved BEC centres. They are aimed primarily at individual learners who wish to obtain a business-related English language qualification and provide an ideal focus for courses in Business English.

The *Business English Certificates* (BEC) were originally developed to meet a specific demand in the Asia-Pacific region. However, the "BEC suite" is now available in over sixty countries worldwide, and it was decided to review the tests to ensure that the needs of a diverse international candidature continue to be fully met. Whilst the tests were under review we took the opportunity to enhance the quality and visual appeal of the papers as well as to improve the reporting of results to give candidates more feedback on the relative strengths and weaknesses of their test performance.

This volume describes UCLES' position on language testing for specific purposes (contexts and uses) as manifested by the BEC and BULATS systems and it sets out clearly the rationale for the changes which were implemented during the BEC revision process.

Cambridge EFL Web-site

Research Notes is not the only source of information about UCLES exams. The Cambridge EFL web-site is another useful source of information.

www.cambridge-efl.org

In this site you will find full information on all of the Cambridge EFL examinations and tests available, as well as the latest news from Cambridge.

The following is a sample of what is available:

Introduction to Cambridge EFL exams

Recognition: Universities and employers which accept Cambridge EFL certificates

The European 5-Level Scale and the work of ALTE

Schools and Centres: Find out where you can take Cambridge EFL exams

News & Updates: UCLES EFL newsletter, *Cambridge First*

Download Publications: Exam handbooks, Sample materials, Past papers

Centre lists

Examiners' reports

Grade statistics

Conferences and Exhibitions and Seminars for Teachers
Forthcoming events and Teacher Seminars sorted by country

Links to related sites:
ALTE, EAQUALS, CEII, Cambridge International Exams (CIE), OCR, etc.

Readers of *Research Notes* may be interested in Grade Statistics and Examiners Reports for the Cambridge EFL examinations.

UCLES EFL Research Papers

For some time now, UCLES EFL has been working to disseminate information about its research, development and validation activities to a wider audience including teachers, applied linguists, language testers and other stakeholders.

A major achievement in this endeavour has been the development of the *Studies in Language Testing* series, published jointly by UCLES and Cambridge University Press for the benefit of test users, language test developers and researchers. The series first appeared in 1995 and since that time twelve volumes have been published; more than half of these report research and validation studies relating directly to the Cambridge EFL examinations. Future publications will include three volumes which chronicle recent major test revision projects – for revised CPE, for revised BEC and for revised CCSE/Oxford (CELS) – see page 23 for more details.

In March 2000 we re-launched *Research Notes*, our newsletter on current developments in research and validation. Subsequent issues appeared in August and November 2000 and in February of this year. *Research Notes* is designed to provide a broad overview of our research activities and to report progress on them as they develop and as results become available. The publication is intended to reach a wide audience of those interested or involved in the Cambridge examinations.

Over the next few months we also hope to start publishing a series of *EFL Research Papers* which will profile specific theoretical and practical issues of interest to us; these papers will also report more fully on the activities we are involved in to explore such issues. From time to time *EFL Research Papers* will include contributions from some of the external consultants and researchers with whom we regularly collaborate.

Two *EFL Research Papers* are in preparation at the present time, both focusing on issues in speaking assessment. The first publication will be a paper reviewing past and current perspectives in the assessment of oral proficiency with particular reference to the Cambridge approach to testing speaking; this will be accompanied by the reprint of a paper on issues in EFL speaking assessment first produced in 1945 by Jack Roach, then a researcher at UCLES. The second publication will be a paper by O'Sullivan, Porter and Weir which was recently commissioned by UCLES EFL to survey the literature on issues in speaking assessment.

EFL Research Papers will be added to the EFL Publications List and will be available for a modest fee directly from UCLES EFL or via the web.

Restructuring within the UCLES EFL Validation Group

Nick Saville, Manager, Research and Validation Group, UCLES EFL

In the first issue of the re-launched *Research Notes* (March 2000), the leading article provided an overview of the research and validation activities being carried out by UCLES EFL (pp2-4). This article provides a brief update with a particular focus on staffing changes which have recently taken place.

Much of the validation work referred to in *Research Notes 1* had been co-ordinated during the 1990s by the Test Development and Validation Group and a list of staff members and their roles was also provided in that issue (*Research Notes 1* p 12). The Group had been responsible for a very wide range of functions including: pre-testing, item banking, institutional testing and other test development projects, as well as a full range of operational research and validation activities. In January 1999 the Performance Testing Unit (PTU) was also established with responsibilities for co-ordinating the Team Leader System for the Speaking tests.

Over the past 18 months the Group has been reorganised and a number of new staff have been recruited. This reorganisation was completed in April 2001 and the Group is now known as the Research and Validation Group.

The Research and Validation Group plays an integral part in all aspects of the UCLES EFL examinations. This includes involvement in the following processes:

- the production of materials and question papers, e.g. item banking, standards fixing, calibration etc.
- the conduct of examinations, including training and evaluation of examiners for both Speaking and Writing
- the grading of examinations and interpretation of results
- the review, evaluation and revision of examinations
- long-term research and development projects, including projects linked to universities

In this role, the Group is expected to provide a service to the other EFL staff who work specifically on developing and producing the examinations and also to provide information for external stakeholders who need to be aware of the quality and fairness of the examinations and to have details of the research and validation work which is carried out.

Members of the Group help in the training of other EFL staff on issues to do with measurement and research design and to write reports and papers for presentation or publication based on the work carried out by the Group. In this capacity the Group organises monthly staff seminars and other training courses and has responsibility for developing and maintaining the reference services provided by the EFL Library. The development of the Cambridge Learner Corpus of written text (with CUP) and of other corpora is also managed by members of the Group.

The Research and Validation work can be categorised into three broad areas of activity:

- Routine Operational Analyses concerning the administration cycle of all exams i.e. exam production, exam conduct, marking/grading, and post-exam evaluation.
- Instrumental Research concerning small-scale projects which are designed to inform the operational activities but which cannot be addressed as part of the routine work (e.g. identified as a requirement in grading or as part of a post-exam review).
- Research Projects concerning long-term research objectives in the field of language assessment which are particularly relevant to our business objectives and future developments.

The Group Manager is responsible to an EFL Research and Validation Steering Group chaired by Director EFL, which oversees and prioritises the work. This work is designed to ensure that all EFL products meet acceptable criteria in relation to the following features of exams:

- Validity
- Reliability
- Impact
- Practicality

In terms of dissemination of information, staff in the Group are responsible for producing *Research Notes*, and a series of *EFL Research Papers* (starting autumn 2001), and they also work with the Deputy Director EFL on the *Studies in Language Testing* (CUP/UCLES) which is now in 14 volumes.

Who's who?

Nick Saville is the Group Manager and overall co-ordinator of the Group. He has been at UCLES EFL since 1989 when the first Evaluation Unit was established, and over the past 10 years he has worked on a wide range of research and development projects. As one of the Assistant Directors in EFL he also has responsibility for co-ordinating a number of other areas, including UCLES EFL activities in countries such as Italy and Mexico for which he is Country Overviewer and the work of UCLES EFL on a range of ALTE projects. **Susan Chapman** is the Group Administrator and assistant to the Group Manager; amongst other things she helps to organise staff seminars and programmes for external visitors.

Neil Jones has been at UCLES since 1992 and is the Senior Research and Validation Co-ordinator in charge of a Unit which is responsible for instrumental research projects with a quantitative focus. These projects include the statistical analysis for the grading and calibration of examinations and in particular the use of Item Response Theory (IRT) in this context. He has also been working on a range of long-term research projects including the development and validation of computer-based tests and the ALTE *Can Do* project. The other staff in the Unit, who have both joined UCLES EFL since the beginning of 2001, are **Ardeshir Geranpayeh** and **Stuart Shaw**.

Nic Underhill is the Senior Research and Validation Co-ordinator in charge of the team conducting the operational analysis of the examinations which is necessary for item banking, grading and post-examination evaluation. He joined UCLES EFL in April 2001 as part of the reorganisation, taking over from Simon Beeston. He is responsible for identifying and documenting the precise requirements of internal customers, producing documentation and analysis for grading sessions, and scheduling the delivery of these requirements on time. He also has a key role in the management of research and validation projects and in the allocation of human resources in the Group as a whole. The other staff in the Unit are **Chris Banks**, **Dave Thighe**, **Helen Marshall**, **Roumen Marinov**, **Tracy Flux** and **Jenny Craft**.

Lynda Taylor worked for many years with UCLES EFL as an independent consultant and was for some time a Chief Examiner for IELTS. She joined UCLES staff in 1999 as the Senior Research and Validation Co-ordinator in charge of the Performance Testing Unit (PTU) which, among other things, manages the training and evaluation programme for oral examiners world-wide. She also helps to co-ordinate the research programme for the performance tests – Speaking and Writing – and other projects which use qualitative research methods. She also has a co-ordinating role with Nick Saville for the overall research programme and for the presentation and publication of the research conducted by UCLES EFL.

The other staff in the Unit are **Janet Bojan**, **Val Sismey** and **Rowena Akinyemi**. **Fiona Ball** has recently joined the Group with a background in Corpus Linguistics. **Chris Hubbard** joined the PTU in July 2001 to support the introduction of the revised Speaking Tests for IELTS, CPE, BEC, and CELS.

Further Information

UCLES provides extensive information on the examinations and assessment services referred to in this newsletter. For further information, visit the UCLES EFL website

www.cambridge-efl.org

or contact

EFL Information
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge CB1 2EU
United Kingdom

Tel: +44 1223 552734
Fax: +44 1223 553068
e-mail: eflinfo@ucles.org.uk

For information on the ALTE five-level scale and the examinations which it covers, visit the ALTE website **www.alte.org**

or contact

The ALTE Secretariat
1 Hills Road
Cambridge CB1 2EU
United Kingdom

Tel: +44 1223 553925
Fax: +44 1223 553036
e-mail: alte@ucles.org.uk

If you would like further copies of this issue of Research Notes, or if you would like to be added to our mailing list (all registered UCLES centres receive copies automatically), please complete this form using block capitals and return it to EFL Information at UCLES. Please photocopy the form if you wish.

☐ Please send me ..... extra copies of this issue of Research Notes.

☐ Please add me to the Research Notes mailing list.

Name .....

Job Title .....

Institution .....

Address .....

.....

.....

Country .....

Please let us have any comments about Research Notes, and what you would like to see in further issues: