3 ResearchNotes

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

NOVEMBER 2000

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

ASSOCIATE MEMBER

ALTE

ResearchNotes

Introduction

Research Notes is the newsletter about current developments in the research, validation and test development work carried out by UCLES EFL.

In this issue, Simon Beeston concludes his series of articles on the UCLES EFL Local Item Banking System with an in-depth look at calibrating items for the IELTS tests. In the last issue of Research Notes, Nick Saville discussed the use of observation checklists to validate tasks for speaking tests – he continues this theme by looking at the development of working checklists as part of the CPE revision project. Lynda Taylor continues the Performance Testing theme, this time focusing on the revision of the performance components.

Issue 3 also introduces many new topics. Lynda Taylor discusses the development of the Public English Test System by the National Examinations Education Authority of China (with support from UCLES EFL) as an example of applying the principles of test development. Simon Beeston looks in detail at the use of statistical analysis.

Computer based testing is one of the fastest growing areas in language testing. Ensuring that computer based and traditional versions of examinations are comparable in terms of difficulty and reliability is an important part of developing computer based versions of certificated examinations. Neil Jones looks at UCLES EFL's research into this area.

UCLES EFL recently introduced new style results slips for FCE, CAE and CPE. Neil Jones discusses how the results slips were developed, and how they help candidates and other test users by giving more information on the skills profile.

In the next issue of Research Notes, there will be articles on reliability, examination revision methodology and a review of the successful entry for the IELTS MA Dissertation Award. We will also be taking a look at work to develop the rating scales for the revised IELTS Speaking Test.

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 UCLES at the address on page 21.

Research Notes is being distributed 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 21.

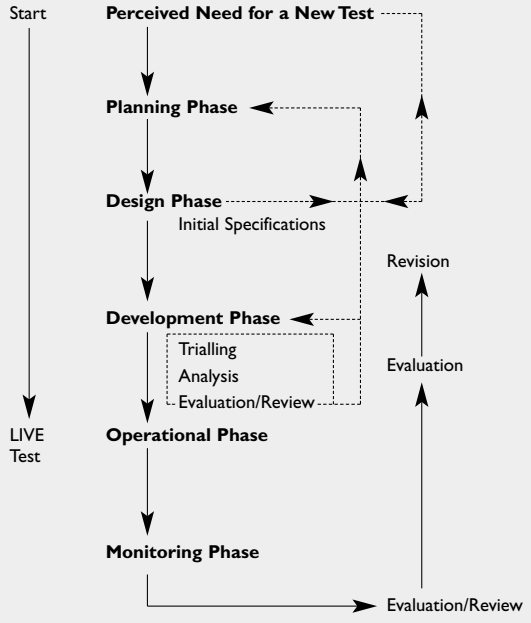
Contents

Published November 2000

Principles and practice in test development: the PETS Project in China

Lynda Taylor, Performance Testing Co-ordinator, UCLES

Issue 1 of Research Notes (March 2000) highlighted a number of different areas of interest for research and validation activity within UCLES EFL. One of these 'strands' focuses on identifying and articulating essential principles that underpin the practice of language test development and revision. Over recent years we have sought to model the complex process of test development as it applies to our own English language examinations, both in the revision of existing tests and also in the development of new tests (see Figure 1). The UCLES model of test development regards the process of test design as cyclical and iterative, in which knowledge and experience gained at different stages are fed back into a process of continuous reassessment.



A Model of the Test Development Process

Figure 1

UCLES EFL is frequently asked to provide consultancy to government ministries or educational institutions which are engaged in language test development projects at national or local level. Since 1995 we have assisted with projects as far afield as Hungary, the Baltic States, Mexico and China. These projects provide a valuable opportunity to refine and validate the UCLES test development model in widely differing assessment contexts. The development from 1997-2000 of the Public English Test System (PETS) in China offered scope for this model to be validated on a large scale and within a very different social and educational context.

The government of the People's Republic of China has always seen proficiency in communicative English as essential to the successful implementation of its Open Door policy in order to encourage rapid modernisation. In 1996 an agreement was signed between China's State Education Commission and the British government's Department for International Development to develop a framework of publicly available English tests beyond the school/college education context – the Public English Test System. The system was designed to provide assessment and certification of communicative English language skills at different levels of competence. The development project was to be managed by the National Educational Examinations Authority (NEEA) in China, with technical assistance from UCLES EFL.

In consultation with NEEA, the essential development principles underpinning work on the PETS system were identified as follows:

- it should take account of the current language teaching and testing situation in China;
- it should provide an adequate focus on communicative language ability;
- it should provide a coherent system of levels linked to one another;
- it should be available and accessible to all learners of the language;
- it should replace current English examinations by virtue of free choice;
- it should be supported by a fully developed and sustainable infrastructure for test delivery, monitoring and ongoing development.

The PETS Project was to comprise three phases covering a three-year period. Phase 1 began in January 1997 with an assessment of the preliminary assessment objectives and criteria. The proposed framework needed to incorporate five distinct but coherent levels, ranging from the level of English expected at Junior High School (after 3 years' English study) to the level required by graduates planning to study and/or work abroad. Draft level criteria, outline test specifications and sample materials were developed by the Chinese test development team with consultancy support from a UK based team at Cambridge. This work was based in part on the UCLES Common Scale Level Criteria, the ALTE Can-Do Statements and the ALTE Level Criteria, all of which needed to be reinterpreted for the Chinese context.

The PETS Development Project		
Phase 1	**Initial Test Design and Development** • Development of level criteria and outline test specifications • production of initial sample materials • feasibility studies	Jan. 1997 – Sept. 1997
Phase 2	**Materials Production and Trialling** • item writer training • test materials production • trialling of materials • analysis and review of results	Sept. 1997 – Nov. 1998
Phase 3	**Live Test Production** • operational test production • live test administration • monitoring of test performance	Nov. 1998 – Jan. 2000

Figure 2

The PETS level criteria describe an overall proficiency scale which is both performance and linguistically oriented. Each of the four skills – listening, reading, writing and speaking – is also described on its own scale. Development of these scales involved careful consideration of their orientation to ensure that they will be appropriately interpreted. User orientation is critical to the transparency of the new system and plays a key role in promotional literature.

The 5-level system is defined in terms of

- a level description,
- formal language knowledge, and
- language use.

This was developed through extensive consultation with examinations officers at NEEA, English language teachers in China, academics in the field of linguistics and pedagogy, and staff at UCLES. The level description defines the likely candidature for each level in terms of age, educational and/or occupational background, etc; the formal language knowledge definition describes the grammatical and vocabulary knowledge expected of candidates at each level; and the language use definition describes what sort of material candidates can handle and what they are expected to be able to do.

The level criteria also include specifications of grammatical content, topic, functions/notions and vocabulary; these are designed to be useful to test-writers, coursebook designers and teachers as well as to the test candidates themselves. Finally, the level criteria give a breakdown of the specific listening, reading, writing and speaking skills considered relevant at each PETS level and which form the testing focus of individual items/tasks within the tests.

The test format for each of the five levels is designed according to a standard template and comprises two separate components: a Written Test contains sections assessing listening comprehension, use of English, reading comprehension and writing; and an Oral Test assesses candidates' speaking ability.

Phase 2 of the project began in September 1997. Further work was done on the level criteria, test specifications and sample materials. Speaking and writing task formats were trialled in China and the results used to develop assessment criteria and rating scales. Item writer guidelines and training materials for all five levels were produced. An extensive infrastructure was established for pretesting and trialling test materials and in June 1998 nearly 5000 students from 22 universities and middle schools all over China participated in pretesting/trialling materials for all test components. This enabled further adjustment of the level criteria and some modification of the test profiles. A comprehensive Test Production Methodology document for the sustainable production of test materials was put in place; and, in a context where candidate numbers for PETS could total more than 5 million annually, extensive procedural and administrative documents were developed for dealing with the issue of oral and writing examiner management.

Phase 3 of the project began in November 1998. In this final phase test-writing, materials editing, pretesting and electronic item-banking activities continued and operational versions of the tests at all five levels were created from the item bank to be ready for the first live test administration scheduled for September 1999. All the necessary systems and supporting documentation were set in place and examiners for the speaking and writing components of the PETS system were identified and trained.

The official launch of the PETS system took place in Beijing in June 1999 and was attended by government officials, NEEA staff, provincial examinations board staff, PETS senior advisers, English teachers, UCLES representatives and around 30 journalists from newspapers, TV and radio. The events of the day were widely reported on national TV, radio and in a selection of national and regional newspapers.

The first live administration of all 5 levels of tests took place as scheduled in September 1999 and proved to be a successful exercise. Over 33,000 candidates sat the tests, most of them doing Levels 1 and 2. PETS Levels 1-4 were administered in cities in 10 of China's 30 provinces while Level 5 tests were held at test centres in 35 universities nationwide. As anticipated, uptake was greatest in the bigger cities and the more developed areas along the coast. The number of candidates entering for PETS was even higher than NEEA had expected, leading them to believe that it will be extremely popular and successful in the future. The plan is now to migrate PETS gradually to other parts of the country according to a predefined plan.

Traditionally, examinations have always played an important social and educational role in China. The promotion of a national public English testing system is of profound social significance in China today for several reasons. It will support the national policy of opening up to the outside world. It will help to address the demand for improvement and development of China's labour force. It will help maintain quality control in a situation where a diversified educational system operates. It will support the concept of life-long learning and ongoing professional development. Finally, it will help integrate the English testing system in China within a coherent and cohesive national framework.

The UCLES model of test development, with its ongoing activity cycle of consultation, drafting, trialling, review and redrafting in all areas of test design, proved to be a sound and effective basis for development work throughout this important international test development project.

The use of Rasch Partial Credit Analysis in test development

Simon Beeston, EFL Validation Manager, UCLES

Previous articles in this series have reported on the processes that UCLES EFL use to ensure that the items it uses in its tests are of the highest possible quality. In this article I am going to look in some detail at a particular type of statistical analysis that reveals the internal functioning of items such as the sentence transformation tasks that are used in the First Certificate in English Use of English paper.

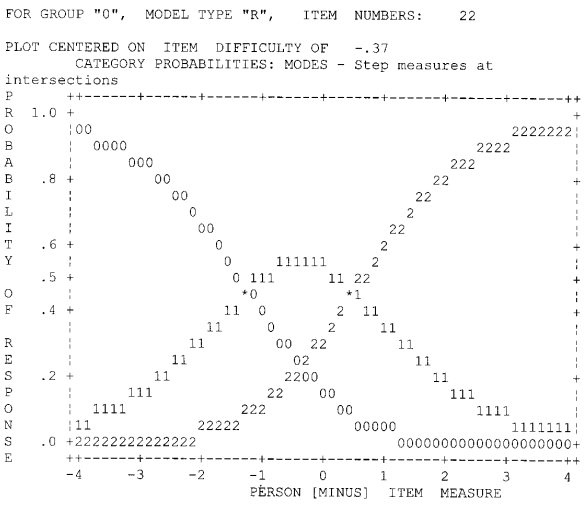
A sentence transformation requires students to rewrite a sentence to produce a new sentence which means the same as the original. Part of the new sentence is provided, along with a key word which must be used to complete the new sentence. An example of this type of item is given below.

> *'Do you know if it will take a long time to develop the film?' the customer asked me.*
>
> **would**
>
> *The customer asked me if I knew how.....*
> *to develop the film.*

The phrase required to complete the sentence is 'if I knew how *long it would take* to develop…' although there may be other possible variations that would be marked correctly. Students successfully completing the sentence would be given two marks; however, if they completed it partially correctly they could still be awarded one mark. For example, an answer including only 'how long' would be awarded one mark for correctly collocating how with long. It is therefore possible to achieve a mark of 0, 1 or 2 depending on the degree of correctness. Items such as these are called scalar items because students can achieve a mark along a scale, albeit a rather short one.

Apart from the issues of item difficulty and item discrimination, it is also important for this type of item that the scale functions properly; that is to say, that for this item, there is a realistic chance of getting 0 or 1 or 2. To determine whether or not this is the case, we use the Rasch Partial Credit model to evaluate how well the scale is working.

Rasch analysis belongs to a branch of test statistics known collectively as Latent Trait Theory. Latent Trait Theory, a term generally used interchangeably with Item Response Theory, arose from dissatisfaction with traditional approaches to educational or psychological measurement. The name 'latent trait' captures the idea that the underlying scores people obtain in a test correspond to a certain amount of ability, not directly observable, but inferable from observing their performance. Furthermore, each test task has a certain amount of difficulty which can be measured relative to the difficulty of other items in the test. Ability and difficulty are mutually-defining terms which allow persons and test items to be ranged along a single unidimensional continuum. In the case of scalar items such as the sentence transformation items reviewed here, the software (Bigsteps; Wright and Linacre, 1994) used by UCLES produces plots showing the probability of each score category being used in relation to the range of ability of the candidates (from low to high ability). This appears as three overlapping distributions (0, 1 and 2) which for a well-functioning, item would produce a roughly symmetrical pattern like the one shown below. The vertical axis provides a measure of probability, the horizontal axis shows increasing ability from left to right. The 0s, 1s and 2s that form the peaks and troughs of the graph show the likelihood of a candidate at a particular ability getting a particular score. As the ability increases, so the likelihood of a higher score increases. The plot below shows that mid-range candidates have a high probability of scoring 1 and that in fact, the scale works well.



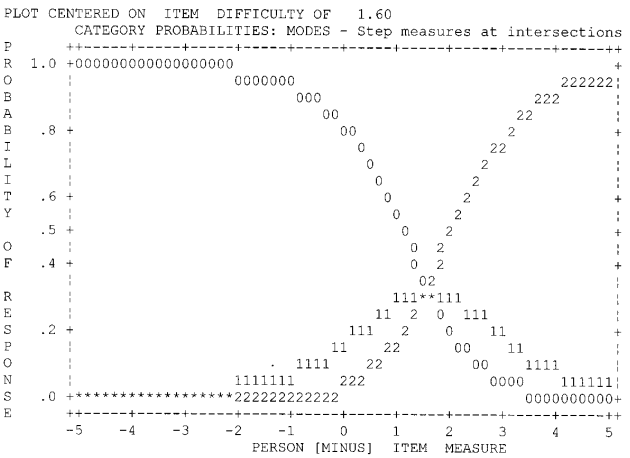
Plot 1

However, one might predict that if the required phrase were something like 'put up with it' where one mark was awarded for 'put up' and another for 'with it', then students would be likely to get all of the phrase or none of it. In such a situation candidates would be more likely to score 0 if they do not know the phrase or 2 if they know all of it. Of course, UCLES item writers endeavour not to produce sentence transformations that function like this but because we pretest all of our material before it is used in a live test, we are able to identify those sentence transformations which do not work as well as they should.

The next example shows a sentence transformation that had been through careful editing but pretesting analysis revealed that the task was not suitable for a live examination.

> *Stephanie lives so far from the school that she has to catch a bus at 6.30 every morning.*
>
> **way**
>
> *Stephanie lives .....from the school that she has to catch a bus at 6.30 every morning.*

The required answer is, of course, '*such a long way*' but pretesting revealed that candidates either knew all of the answer or none of it. Compare the distributions of probable scores in the plot below with the earlier well-functioning plot.



Plot 2

As can be seen from this second plot of probable scores, the likelihood of scoring 1 is always lower than that of scoring either 0 or 2. Instead of being a scalar item, it effectively becomes a dichotomous item that is double weighted.

UCLES EFL use a number of different quality control stages to ensure that all material used in live examinations is of the highest possible quality. Using pretesting statistics to evaluate how scalar items are functioning is one such quality control stage in a process that is designed to ensure our examinations are fair, reliable and useful for all our candidates.

References:

Wright, B. D. & Linacre, J. M (1994): *BIGSTEPS*, Mesa Press

Developing observation checklists for speaking-tests

Nick Saville, Manager, EFL Test Development and Validation Group, UCLES

Barry O'Sullivan, University of Reading

In the last issue of Research Notes, we discussed the background to developing and using observation checklists to validate speaking tests. This article looks at the development of the working checklists through the collaboration of UCLES staff with the Testing and Evaluation Unit at Reading University, including Don Porter, Barry O'Sullivan and Cyril Weir.

Weir (1993), building on the earlier work of Bygate (1988), suggests that the language of a speaking test can be described in terms of the informational and interactional functions and those of interaction management generated by the participants involved. With this as a starting point, the group of researchers at the University of Reading were commissioned by UCLES EFL, to examine the spoken language, Second Language Acquisition (SLA) and language testing literatures to come up with a initial set of such functions. These were then presented as a draft set of three checklists representing each of the elements of Weir's categorisation, as set out in figure 1.

The next concern was to develop a working version of the checklists to be followed by an evaluation of using this type of instrument in real time, using either tapes or live speaking tests. The concept that drives the development model (figure 2) is the evaluation at each level by different stakeholders. At this stage of the project these stakeholders were identified as:

- The consulting expert testers (the University of Reading group)
- The CPE Revision Project Team (including the Materials Development and Validation staff at UCLES)
- UCLES Senior Team Leaders (i.e. key staff in the oral examiner training system)

All these individuals participated in the application of each draft.

In order to arrive at a working version of the checklists, a number of developmental phases were anticipated. At each phase, the latest version (or draft) of the instruments was applied and this application evaluated.

Informational functions	
Providing personal information	give information on present circumstances give information on past experiences give information on future plans
Providing non-personal information	give information which does not relate to the individual
Elaborating	elaborate on an idea
Expressing opinions	express opinions
Justifying opinions	express reasons for assertions s/he has made
Comparing	compare things/people/events
Complaining	complain about something
Speculating	hypothesise or speculate
Analysing	separate out the parts of an issue
Making excuses	make excuses
Explaining	explain anything
Narrating	describe a sequence of events
Paraphrasing	paraphrase something
Summarising	summarise what s/he has said
Suggesting	suggest a particular idea
Expressing preferences	express preferences
Interactional functions	
Challenging	challenge assertions made by another speaker
(Dis)agreeing	indicate (dis)agreement with what another speaker says (apart from 'yeah'/'no' or simply nodding)
Justifying/Providing support	offer justification or support for a comment made by another speaker
Qualifying	modify arguments or comments
Asking for opinions	ask for opinions
Persuading	attempt to persuade another person
Asking for information	ask for information
Conversational repair	repair breakdowns in interaction
Negotiating meaning	check understanding attempt to establish common ground or strategy respond to requests for clarification ask for clarification make corrections indicate purpose indicate understanding/uncertainty
Managing Interaction	
Initiating	start any interactions
Changing	take the opportunity to change the topic
Reciprocity	share the responsibility for developing the interaction
Deciding	come to a decision
Terminating	decide when the discussion should stop

Figure 1

Phase 1

The first attempt to examine how the draft checklists would be viewed and applied by a group of language teachers was conducted by Angela ffrench (1999) from the UCLES EFL team working on the CPE Revision. The group of Teachers were attending a seminar in Greece. Of the participants at the seminar, approximately 50% of the group reported that English (British/American/Australian) was their first language, while the remaining 50% were native Greek speakers.

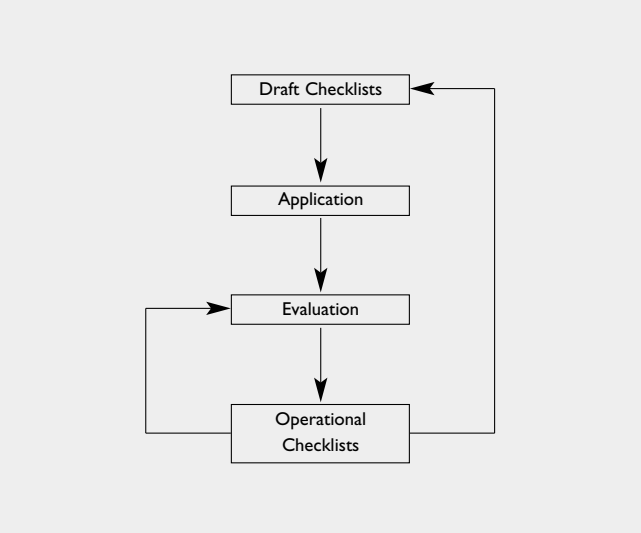


Figure 2: The Development Model

In their introduction to the application of the Observation Checklists (OCs), the participants were given a series of activities which focused on the nature and use of those functions of language seen by task designers at UCLES to be particularly applicable to their EFL Main Suite Speaking Tests (principally FCE, CAE and CPE). Once familiar with the nature of the functions (and where they might occur in a test), the participants applied the OCs in 'real' time to an FCE Speaking Test from the 1998 Standardisation Video. An FCE video was used as one of the objectives of the CPE Revision project was to bring the Speaking Test format in line with the other Main Suite examinations.

Of the 37 participants, 32 completed the task successfully – that is they attempted to make frequency counts of the items represented in the Observation Checklists, although there was some disagreement as to the frequency of the use of language functions. However, when the data was examined from the perspective of agreement on whether a particular function was observed or not (ignoring the count, which in retrospect, was highly ambitious considering the lack of systematic training in the use of the questionnaires given to the teachers who attended), we find that there is a striking degree of agreement on all bar a small number of functions (figure 3).

This aspect of the developmental process was considered to be quite successful. At this stage it was felt that the rarely checked items may not be the most serious problem (as these may represent observational errors). Instead the middle range of items appear to have caused a greater degree of confusion, and so are marked for further investigation.

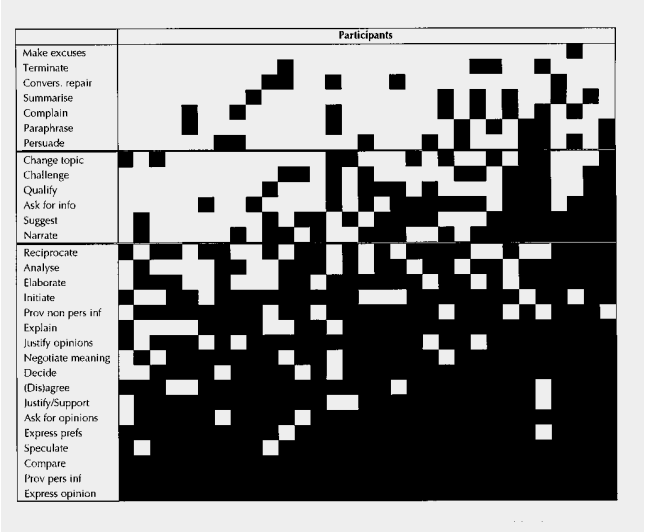


Figure 3 In order to make these patterns of behaviour clear, the data have been sorted both horizontally and vertically by the total number of observations made by each participant and of each item. A number of elements within the scale caused some difficulty. These are highlighted by the bold lines. Items above the lines have been checked by some participants, in one case by a single person, while those below the line have been checked by a majority of participants.

Phase 2

In this phase, a much smaller gathering was organised, this time involving members of the development team as well as the three UK-based UCLES Senior Team Leaders. In advance of this meeting all participants were asked to study the existing checklists, and to exemplify each function with examples drawn from their experiences of the various UCLES main suite examinations. The resulting data were collated and presented as a single document that formed the basis of discussion during a day-long session. Participants were not made aware of the findings from Phase 1.

During this session many questions were asked of all aspects of the checklist, and a more streamlined version of the three sections was suggested. Many of the problem items identified in Phase 1 were dealt with. Some were omitted from the checklists, included in a different category or given more clarification. From this, a working version of the checklists was produced for use in the next phase.

Phase 3

In the third phase, the revised checklists (figure 4) were given to a group of 15 MA TEFL students who were asked to apply them to two FCE tests (both involving a mixed-sex pair of learners, one pair of approximately average ability and the other pair above average). Before using the checklists, the group had a pre-session task to familiarise themselves with FCE format and tasks.

Informational Functions	
Providing personal information	give information on present circumstances give information on past experiences give information on future plans
Expressing opinions	express opinions
Elaborating	elaborate on, or modify an opinion
Justifying opinions	express reasons for assertions s/he has made
Comparing	compare things/people/events
Speculating	speculate
Staging	separate out or interpret the parts of an issue
Describing	describe a sequence of events describe a scene
Summarising	summarise what s/he has said
Suggesting	suggest a particular idea
Expressing preferences	express preferences
Interactional Functions	
Agreeing	agree with an assertion made by another speaker (apart from 'yeah' or non-verbal)
Disagreeing	disagree with what another speaker says (apart from 'no' or non-verbal)
Modifying	modify arguments or comments made by other speaker or by the test-taker in response to another speaker
Asking for opinions	ask for opinions
Persuading	attempt to persuade another person
Asking for information	ask for information
Conversational repair	repair breakdowns in interaction
Negotiating meaning	check understanding indicate understanding of point made by partner establish common ground/ purpose or strategy ask for clarification when an utterance is misheard or misinterpreted correct an utterance made by other speaker which is perceived to be incorrect or inaccurate respond to requests for clarification
Managing Interaction	
Initiating	start any interactions
Changing	take the opportunity to change the topic
Reciprocating	share the responsibility for developing the interaction
Deciding	come to a decision

Figure 4

Prior to the actual session, the group was given an opportunity to have a practice run using a third FCE examination. While this training period, coupled with the pre-session task, was intended to provide the students with the background they needed to consistently apply the checklists, there was a problem during the session itself. This problem was caused by the

failure of a number of students to note the change from Task 3 to Task 4 in the first test observed – possibly caused by a lack of awareness of the test structure itself and not helped by the seamless way in which the examiner on the video moved from a two-way discussion involving the test-takers to a three-way discussion. This meant that a full set of data exists only for the first two tasks of this test. As the problem was spotted in time, the second test caused no such problems. The participants were asked to record each function when it was first observed, as it was felt that without extensive training it would be far too difficult to fully apply the OCs in real time.

Phase 4

In this phase a transcription was made of the second of the two interviews used in Phase 3 – as there was a full set of data available for this interview. The OCs were then mapped on to this transcript, to give an overview from a different perspective of what functions were generated – it being felt that this map would result in an accurate description of the test in terms of the items included in the OCs.

Finally, the results of Phases 2 & 3 were compared. This indicated that the checklists were working quite well, although there were still some problems in some items. Feedback from participants suggested that this may have been due to misunderstandings or misinterpretations of the gloss and exemplifications used. These issues were resolved in the later phases of the Revision Project, with version of the checklists being piloted in trials.

Discussion and Conclusions

While this article has focused on the validation procedures, the checklists are also relevant to task design. By taking into account the expected response of a task (and by describing that responses in terms of these functions) it is possible to explore predicted and actual test-task outcome. This is a useful guide for item writers in taking a priori decisions about content coverage. Through this approach it should be possible to predict linguistic response more accurately, in terms of the elements of the checklists, and to apply this to the design of test-tasks – and of course to evaluate the success of the prediction later. In the longer term this will lead to a greater understanding of how tasks and task formats can be manipulated to result in specific language use. The checklists do not make it possible to predict language use at a micro level (grammatical form or lexical), but they do enable test-developers to predict informational and interactional functions and features of interaction management – a notion supported by Bygate (1999).

The checklists also enable us to explore how systematic variation in such areas as interviewer questioning behaviour (and interlocutor frame adherence) affects the language produced in this type of test. For example, in the interview transcribed for this study the examiner directed his questions very deliberately (systematically aiming the questions at one participant and then the other). This may have had the effect of limiting spontaneity in the intended three-way discussion in the test, so occurrences of Interactional and Discourse Management Functions did not materialise to the extent intended by the task designers. This also raises implications for the way examiners are trained to manage this part of the test.

The checklists require a degree of training and practice similar to that given to raters if a reliable and consistent outcome is to be expected. To achieve this, standardised training materials for Oral Examiners were developed alongside the checklists.

The potential of the checklists as an evaluative tool is great: it is hoped it can address and provide more comprehensive insight into issues such as:

- The language functions the different task-types employed in the UCLES main suite Paper 5 (Speaking) typically elicit
- The language which the paired-format elicits and how it differs in nature and quality from that elicited by interlocutor-single candidate testing

In addition to these issues, the way in which the checklists can be applied may allow for other important questions to be answered. For example, by allowing the evaluator multiple observations (stopping and starting a recording of a test at will), it will be possible to establish whether there are quantifiable differences in the language functions generated by the different tasks – i.e. the evaluators will have the time they need to make frequency counts of the functions.

References

Bygate, M (1988): *Speaking*, Oxford: Oxford University Press.

Bygate, M (1999): Quality of language and purpose of task: patterns of learners' language on two oral communication tasks, *Language Teaching Research*, 3(3), 185-214.

ffrench, A (1999): *Language Functions and UCLES Speaking Tests*, Seminar in Athens, Greece, October 1999.

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

Acknowledgements

We would like to thank Cyril Weir and Don Porter for input into the first version of the checklist. Significant input was also received from members of the CPE Revision team (including Angela ffrench, Lynda Taylor, Anne Gutch and Cristina Rimini) and a group of UCLES Senior Team Leaders.

BULATS: A case study comparing computer based and paper-and-pencil tests

Neil Jones, Research Co-ordinator, UCLES EFL

Introduction

The growth of computer based testing

Computer based (CB) testing is a relatively recent development in UCLES EFL and in many ways is still in a developmental stage. Compared with the large candidatures for the major paper-and-pencil (P&P) exams the current market for CB products is generally associated with low-stakes testing: they are not certificated in the same way as the main suite exams, the conditions in which they are administered are not supervised by UCLES, and they are shorter.

However, in the future this situation will change. The administration of certificated exams, probably online, is a possibility and is an area of current research. A CB version of IELTS has been trialled, and will be made available as an alternative format in 2001.

Current CB products produced by UCLES EFL, in partnership with ALTE members, include:

- BULATS, a CB alternative to the P&P BULATS (Business Language Testing Service), available in English and French;
- Linguaskill, a computer adaptive test (CAT) with a business focus, developed for Manpower Europe and now available in English, French, German, Spanish and Dutch;
- Placement tests under development for the British Council and OUP.

Comparing CB and P&P tests

All UCLES EFL tests and exams provide results which can be interpreted in terms of ALTE levels. Thus there is a general requirement to ensure that in terms of level, there is comparability across all products, CB and P&P. This is also true of different language versions of multilingual CATs like Linguaskill. A major area of research, which is particularly important for establishing this kind of comparability, is the development and use of 'Can-do' statements to provide a basis for defining levels in functional terms. An update on this project was reported in *Research Notes* No. 2.

Every test contains measurement error, and is subject to practical constraints such as length, range of skills tested, etc. In comparing CB and P&P formats, it is important to distinguish general issues of reliability and test relatedness, which affect any comparison between tests, from specific issues relating to the testing format.

Specific issues in the comparison of CB and P&P formats include:

1. The difficulty of particular task types and of individual items;
2. The overall level of performance, and the spread of scores;
3. The impact of such features of CB test administration as time limits, enabling or disabling review of earlier responses, etc;
4. The effect of such test-taker features as gender, age, or familiarity with computers, both individually and when grouped e.g. by country of origin, professional background etc.

These are relevant to the comparison of CB and P&P formats of a linear test such as IELTS.

Additionally, where the CB test is adaptive, (e.g. Linguaskill, BULATS, and the OUP and British Council Placement Tests) the following issues arise:

1. The effect of an adaptive mode of administration on test reliability, discrimination and the effective scale length of the CAT format;
2. The effect of guessing in the P&P format.

This paper focuses on a particular project which was recently completed: a comparison of the CB and P&P forms of BULATS, which addresses several of the issues listed above.

Item banking: the basis of comparability

It is important to understand that when we compare scores between CB and P&P formats we generally do not mean raw scores. Most current CB products are adaptive tests. In such a test candidates will tend to achieve roughly similar proportion-correct scores. But clearly a candidate who scores 60% on a set of difficult items has demonstrated more ability than the candidate who scores 60% on an easy set of items. The scores we are talking about are actually ability estimates derived from a latent trait (Rasch) analysis (see Simon Beeston's article on p 4 for an introduction to Rasch measurement). Similarly, the raw scores on the P&P version are Rasch-analysed to derive ability estimates. It is these which we can compare.

To estimate ability using Rasch techniques we must first know the difficulty of each item in the test, and a basic condition for constructing comparable tests is that the items used in both should be taken from a pool, or item bank, of items which have been calibrated (their difficulty estimated) on the same scale. UCLES EFL has for some years been using item banking techniques in the routine test construction cycle, so that generally when items are made available for use in a CB test their difficulty is known with some precision.

The BULATS comparability study

Earlier this year, 85 learners of English agreed to do a CB and P&P version of the BULATS test of English. They also completed a questionnaire.

Findings from the questionnaire

Candidates were asked to say:

1. How difficult they found the two forms of test;
2. Whether they found the two forms of test to be of appropriate length;
3. Whether they liked using computers;
4. Which version of the test they liked best;
5. Whether they considered themselves good at using computers.

The questionnaire produced some interesting findings, but no evidence that personal attitudes to computers affected performance on the test.

Most people said they liked using PCs. More than half preferred the CB version, and there was a clear tendency for people who preferred the P&P version to say that they found this version easier than the CB version. There was also a small tendency for people who claimed not to be good at using computers to say they found the P&P test easy, but the CB test hard. However, there was no relation between any of these statements and the final score in either form of the test.

These findings suggest that for this group of subjects, who were studying in Cambridge when they took the tests, there was no effect on scores connected with computer familiarity, like or dislike. This in turn suggests that typical BULATS candidates would most probably not be disadvantaged or advantaged by such factors.

Reliability of each test

For the P&P test, classical Alpha and Rasch estimates of reliability were .93 and .92 respectively. An average reliability was estimated for the CB tests of .94. Thus both these tests show good reliability for this sample of respondents.

These reliability estimates are based on internal consistency estimates. It would be useful to have coefficients of stability from test-retest data. These could be directly compared with the correlations found between CB and P&P formats, and would thus indicate whether differences in test format have a significant effect on correlations. However, in the absence of test-retest data we can use the square of the alpha reliability to model the correlation between two sittings of the test. This gives .88 for the CB format and .87 for P&P.

Correlation between CB and P&P scores

Figure 1 shows a scatterplot of the CB and P&P scores. The correlation before outliers are removed is .77. With 6 outlying cases removed it is .86. Inevitably the experimental conditions, where the two tests were completed one after the other, produce variations in performance due to fatigue, inattention etc. Removing a small number of outlying cases is sufficient to produce an actual correlation between test formats which is similar to the modelled test-retest reliability for each test format taken separately (as presented above). While actual test-retest data will allow us to settle this question with more confidence, it appears that the effect of test format on the correlation of test results was minimal for this group of respondents.

Overall level and spread of scores

Figure 1 indicates (from the way the points are distributed along the identity line) that there is good agreement in overall level between the scores obtained on the two formats. However, the spread of scores is clearly narrower for the P&P format, as indicated by the slope of the trend

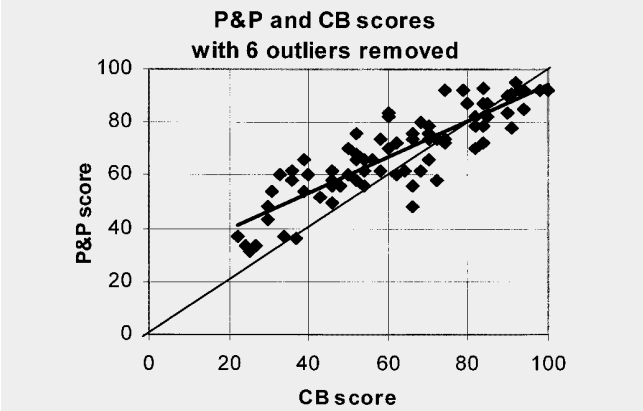


Figure 1 BULATS CB and P&P ability scores compared

line which has been added. In other words the CB test format is slightly more discriminating. The linear trend describes the relationship well: curvilinear trends (e.g. 2 or 3-order polynomials) do not account for significantly more common variance.

	CB	P&P
mean	2.80	3.11
SD	1.24	1.15

Table 1 Mean and SD of scores on CB and P&P test formats

Table 1 shows the mean and SD of scores on both formats. The P&P scores are higher overall, and this is mostly caused by lower ability candidates performing better on the P&P version.

The narrower spread of scores on the P&P version of a test has been observed previously in other contexts, and is characteristic. The adaptive CB test selects the most appropriate items for each candidate, according to their estimated level. It gives each candidate a chance to show just how high or low their level is. The P&P test is the same for all candidates, and necessarily each item gives slightly less information, because it is of inappropriate level for a proportion of the candidates.

The effect of 'guessing'

It appears that a crucial aspect of this difference between CB and P&P is what is commonly called guessing, although this is better characterised as the contribution of chance in a response to an item. While there is no systematic benefit from guessing in an adaptive test format, the P&P format does enable candidates who guess to score higher (under normal scoring rules where wrong answers are not penalised).

Guessing is an unfortunate label, because it suggests a distinct, aberrant and relatively rare type of behaviour which occurs only when a respondent finds an item to be wholly too difficult. In fact, what we call guessing is not a distinct type of behaviour, but just the extreme end of a continuum, where the relative contributions of chance and ability are 100% and zero respectively.

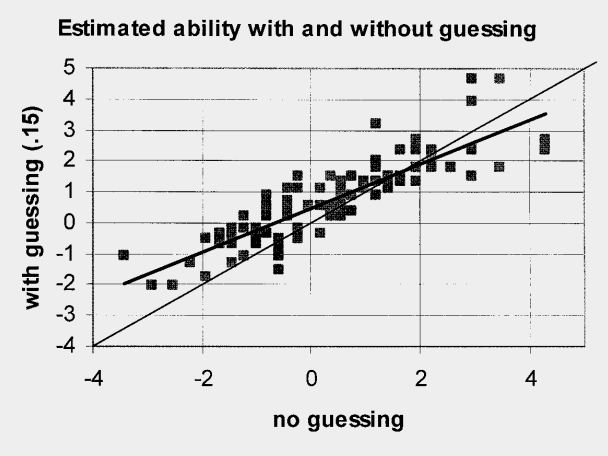


Figure 2 The effect of guessing on ability estimates (from simulated data)

Simulated response data allow us to examine the effect of chance on ability estimates. Figure 2 shows a scatterplot comparing estimates of ability from two artificially-generated datasets. Both sets were generated from the same set of abilities and difficulties. The first used the standard Rasch model; the second used a modified model in which the probability of a correct response tends to be an arbitrary lower limit of 15%.

There is a striking resemblance between this figure and Figure 1 – the comparison of scores on CB and P&P versions of BULATS. What is particularly interesting is that, as with the CB – P&P score comparison, the trend line plotted through the data points is linear. A more complex curvilinear relationship accounts for no more of the common variance. This shows that the effect of chance is not limited to lower-ability candidates, but affects ability estimates proportionately across the whole scale.

Discussion

This paper has not addressed all the issues relevant to the comparison of CB and P&P test formats. The comparability of test content, and the performance of particular task types, have not been treated. However, the findings of the BULATS comparability study described here have contributed significantly to our understanding of how CB and P&P test formats relate, and support a view that it should be practical to develop the two formats for use interchangeably.

Each test format was found to be highly reliable for this group of subjects. The correlation between scores on the two tests was high, and removing just a small number of cases of poor agreement was sufficient to produce a correlation as high as the theoretical (squared alpha) test-retest correlation of each test format taken separately. In other words, there was no evidence of the test format having an important effect on the correlation between two attempts at the test.

The questionnaire also showed no relationship between attitude to computers and test scores on the CB test format, for this group of respondents. Thus on this evidence the two forms of test appear to measure the same thing; however, they clearly measure it on a different scale, as shown by the narrower score range observed for the P&P test format.

The relation between CB and P&P scores was found to be linear. A study conducted on generated response data confirmed that a similar linear relationship could be produced by modelling the effect of chance, or 'guessing', which affects P&P scores much more than CB scores. Thus there is a theoretical explanation for the difference in the observed score distributions, and so it should be possible to equate scores on the two test formats by a suitable linear scaling.

The comparability of CB and P&P formats is practically of great importance. Clearly, decisions on how to report the equivalence of different test formats require consideration of such issues as how high-stakes the test is, who the users of the test are, and whether a simple form of report is practically more useful than a psychometrically rigorous but less transparent one. In the case of BULATS, it seems both reasonable and useful to aim at using a single scale to report scores on both computer-based and paper-based forms of the test.

EASA 2000

We are pleased to announce that UCLES EFL has won the European Academic Software Award (EASA) for CommuniCAT.

CommuniCAT is the multilingual, computer adaptive language testing engine that drives such UCLES EFL products as CB BULATS, the British Council Placement Test and the UCLES/OUP Quick Placement Test. It has also been developed in a range of European languages by members of the ALTE sub-group known as KoBALT.

The EASA competition is held biennially under the auspices of EKMA, the European Knowledge Media Association. The EASA 2000 competition was held in Rotterdam. This time there were 235 entries from all over Europe, out of which 30 were selected as finalists by an evaluation involving students, users, teachers and software experts drawn widely from European countries.

An international team of 21 Jurors made the final selection of 10 Award Winners during an intensive three day meeting, culminating in the Award Ceremony itself at the World Trade Centre, Rotterdam, on Tuesday 28th November.

Sarah Corcoran represented EFL over the three day event, and accepted the Award from the Dutch Minister of Education, Loek Hermans.

The CBT Team, co-ordinated by Michael Milanovic (Deputy Director EFL), has drawn on the skills and contributions from many in UCLES EFL over the past five years during which CommuniCAT has been developed. The work on item banking and the calibration of items has been particularly important and the research of Neil Jones from the Validation Group into computer adaptive testing was particularly commended by the EASA Jury.

For more information on BULATS, please visit the BULATS website – **www.bulats.org**. Issue 9 of *Cambridge First*, available from EFL Information, contains articles about the OUP and British Council placement tests.

Approaches to rating scale revision

Lynda Taylor, Performance Testing Co-ordinator, UCLES

Whenever UCLES undertakes to revise an existing examination, the special needs of the performance testing components within that examination (i.e. speaking and writing tests) have to be carefully considered.

The overall methodology which UCLES EFL employs in test revision projects was outlined in the previous issue of *Research Notes* (August 2000). Revision of the performance testing components is especially complex because it usually involves a wide range of factors including redesign of the test format, redevelopment of the criteria for assessment, revision of the measurement scales and redrafting of the performance descriptors to be applied by the raters or examiners; considerable resources will also be required to develop appropriate methods and materials for retraining and standardising writing and speaking examiners in readiness for the revised test becoming live. Issue 1 of *Research Notes* (March 2000) highlighted some of the research issues relating to the nature of rating scales in speaking assessment; many of these issues relate to rating scales for writing assessment as well.

Traditionally the design and construction of rating scales for direct tests of writing and speaking ability have depended upon an a priori approach; in this approach assessment criteria and rating scale descriptors are developed by 'experts' (i.e. teachers, applied linguists and language testers) using their own intuitive judgement. In recent years several writers in these fields have advocated a more empirically-based approach to rating scale construction (Shohamy, 1990; Upshur and Turner, 1995; Milanovic, Saville, Pollitt and Cook, 1996; Fulcher, 1996). An empirically-based approach involves analysing samples of actual language performance in order to construct (or reconstruct) assessment criteria and rating scale descriptors; it also involves investigating the way in which these are likely to be interpreted and applied by human raters.

In practice, we are rarely able to approach the task of rating scale development with a blank sheet of paper before us; this is usually only possible when developing a brand new test. Revision of an established test, on the other hand, usually involves redeveloping the existing criteria and scales to take account of advances in applied linguistics, pedagogy, testing and measurement theory. The process normally begins with a review of the historical development of the present rating scales to consider their original rationale and orientation and to evaluate their strengths and weaknesses.

Through the analyses of test score data carried out routinely after each test administration we can monitor over time how the assessment scales are functioning and can identify possible problems to do with scale length, scale integrity, and scale interpretation. Conversational and discourse analytic techniques help us to investigate samples of speaking test interviews or writing test scripts at different proficiency levels and to confirm criterial features of test-taker performance. Finally, qualitative feedback gathered from examiners on their experience of using the rating scales helps to inform plans for redevelopment, as does the theoretical and practical experience gained from revision projects for our other UCLES EFL examinations.

Once the initial comprehensive review is complete, the process moves on to the drafting of revised assessment criteria and rating scale descriptors. These revised criteria and descriptors will go through successive cycles of trialling and redrafting using a variety of different approaches.

For example, multiple rating exercises – in which a team of experienced examiners applies the draft scales to candidate performances – enable us to carry out analyses using multi-faceted Rasch (FACETS) and generalizability theory (GENOVA); this means we can investigate questions such as:

- Do the scales measure distinct aspects of language proficiency?
- Do they contribute consistently to the candidate's final score?
- Do raters use and interpret the markscheme in the same way?
- Do candidates score in the same range on the current and revised rating schemes?

Conversational and discourse analytic studies of sample writing and speaking performances at different proficiency levels help us to answer questions such as:

- What are the features of language which distinguish different levels of performance?
- Is the revised task design capable of eliciting a broad enough sample of candidate output against the revised assessment criteria and rating scales?

Finally, we use focus group techniques and verbal protocol analysis with examiners as they actually apply the draft criteria and scales to sample performances; this provides us with additional insights into the theoretical and practical problems they encounter, e.g.

- What do raters pay attention to in their rating?
- How do raters reach a final decision in their rating?
- Do raters find certain criteria more difficult to identify and scale than others?

Answers to these questions are especially valuable in informing our development of materials for rater training and standardisation.

Alan Tonkyn (1999) has identified the essential qualities of assessment scales as:

- theoretical relevance
- discriminating power
- assessability

This is consistent with UCLES' commitment to balancing the qualities of validity, reliability, impact and practicality in test design/use. When redesigning and improving the assessment criteria and rating scales for our tests we seek to achieve an optimum balance among these four qualities.

UCLES EFL is currently focusing considerable resources on a number of important revision projects: CPE Writing and Speaking, BEC Writing and Speaking, CELS (formerly CCSE/Oxford) Speaking and Writing, and the IELTS Speaking Test. For all these projects the redevelopment of assessment criteria and rating scales depends on a data-driven rather than a purely intuitive approach, but is still supplemented with insights derived from expert judgement. By combining the use of quantitative and qualitative methodologies it is possible to redevelop valid, reliable and practical assessment criteria and rating scale descriptors; such studies also feed directly into the development of strategies for retraining and standardising examiners. The next issue of *Research Notes* will report in detail on the project to redevelop the assessment criteria, rating scale and band descriptors for the revised IELTS Speaking Test, scheduled for introduction in July 2001.

References and further reading

Fulcher, G (1996): Does thick description lead to smart tests? A data-based approach to rating scale construction, *Language Testing*, Volume 13/2, pp 208-238

Milanovic, M, Saville, N, Pollitt, A and Cook, A (1996): Developing rating scales for CASE: theoretical concerns and analyses, in Cumming, A and Berwick, R (Eds) *Validation in Language Testing*, Multilingual Matters, Clevedon, pp 15-33

Shohamy, E (1990): Discourse analysis in language testing, *Annual Review of Applied Linguistics*, Volume 11, pp 115-128

Tonkyn, A (1999): *Reading University/UCLES IELTS Rating Research Project* – Interim Report.

Upshur, J A and Turner, C E (1995): Constructing rating scales for second language tests, *English Language Teaching Journal*, Volume 49/1, pp 3-12

New-style statements of results

Neil Jones, Research Co-ordinator, UCLES

As part of our efforts to improve the reporting of examination results and provide more useful feedback, UCLES is gradually introducing new-style 'Statements of Results' for FCE, CAE and CPE from mid-2000 and for KET and PET from early 2001. Other examinations will follow later.

The following explanatory notes have been issued to accompany the new-style result slips.

Every candidate will be provided with a Statement of Results which includes a graphical display of the candidate's performance in each component. These are shown against the scale Exceptional – Good – Borderline – Weak and indicate the candidate's relative performance in each paper.

In looking at this graphical display it is important to remember that the candidates are NOT required to reach a specific level in any component, i.e. there are NO pass/fail levels in individual components. Thus different strengths and weaknesses may add up to the same overall result.

We recommend that fail candidates planning to resit an examination, or pass candidates who plan to continue their studies, do not focus only on those areas where they have a performance which is less than Borderline, but try to improve their general level of English across all language skills.

The profile indicates a candidate's performance on the specific occasion when they sat the exam – this may be influenced by a number of different factors, and candidates can find that they have a somewhat different profile on another occasion. Evidence of candidates who resit exams indicates that in some cases performance declines overall and in other cases declines in some papers while improving in others.

The information on these new-style Statements of Results replaces the indications of High Performance/Particularly Weak Performance provided previously.

This paper looks at the interpretation of these new-style statements in the context of a discussion of the UCLES EFL approach to grading.

Let us begin by considering the construct of English language proficiency which the exams operationalise. Essentially, the exams take an inclusive view: the fairest measure of proficiency reflects a candidate's aggregate performance over the whole range of language skills. That is why the UCLES EFL exams have up to five component papers.


Statement of Results

ENGLISH AS A FOREIGN LANGUAGE

First Certificate in English

Candidate Profile

Explanatory Notes

UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

New style statement of results

This approach originates in the way most UCLES EFL exams fit into a pedagogical process, reflecting all aspects of language study and providing positive feedback into the teaching and examination preparation cycle. It is also in line with modern theories of communicative language ability, to the extent that these present a complex picture of interdependent competences defying a reductionist psychometric approach.

This view of language proficiency is reflected in the approach to grading. It is the examination as a whole which is failed or passed: there are no hurdles in individual papers. Thus candidates can and do achieve a passing mark in very different ways, representing quite varied profiles of skills.

The details of scoring vary across exams, but the general picture is as follows. Papers contain different numbers of items, but the marks are usually equally weighted. FCE, for example, has five papers weighted to 40 marks each, the marks being summed to achieve an examination score out of 200. Mark distributions are not scaled to have an equal standard deviation (although linear scaling may be applied to adjust for rater severity in the case of the Writing paper, or differential version difficulty in the case of the Listening paper). The papers are graded in such a way that the marks indicating a satisfactory level of performance in each paper sum to a passing grade in the examination.

This begs the question of how criterion levels of performance are defined for each level of UCLES EFL examination. Are criterion standards fixed for each paper, and if so how?

The UCLES EFL approach has both normative and criterion-related features. In criterion terms, each examination level can be seen as representing a level of proficiency characterised by particular abilities to use English to some purpose. Exams at ALTE Levels 1 and 2 were designed from the outset around the Council of Europe Waystage and Threshold level specifications, which have an essentially functional definition. At Level 3 the direction has been reversed, with the Council of Europe Vantage Level being constructed to describe the features of this level, already well established in English language teaching and in the Cambridge FCE examination. The development of detailed functional descriptions of each ALTE/Cambridge level (the Can Do Project) is currently the focus of a significant research effort. Thus increasingly informative real-world criteria for interpreting the meaning of UCLES exams are becoming available for all stakeholders in the testing process.

The normative aspect relates to the way that the target difficulty of each component paper is set, with the aim of making each paper in an examination of similar difficulty for the typical candidate. A facility level of about 60 per cent is the test construction target for exams at the three upper levels: this should indicate a satisfactory level of performance if repeated across all papers. Typical levels of performance in the different skills reflect, of course, the makeup of the global candidature for a given examination. This normative aspect of the UCLES EFL approach is captured in two different ways. For the objectively-marked papers, it is specified in the test design as a mean item difficulty (a Rasch modelling approach). Subjectively-marked papers (the Writing and Speaking papers) depend on the application of a rating scale by trained and standardised raters. The aim of the rating scale and the associated training effort is to make satisfactory performance equate to a score of about 60 per cent, as with the objective papers.

While the users of the UCLES EFL main suite exams (KET, PET, FCE, CAE, CPE) are still overwhelmingly in favour of the current approach to grading, with a single examination grade, there is at the same time a demand for more information concerning the way the grade was arrived at. This reflects the pedagogical context in which UCLES EFL exams are generally taken – feedback on performance in each paper is seen as a guide for further study, particularly in the case of failing candidates who wish to re-take the examination.

The purpose of the new profiled result slips is to give useful information about performance in each paper. What are plotted in the result slips are not candidates' raw marks, but marks which are scaled to implement the normative frame of reference which has been presented above. The candidate with a borderline pass, if his/her skills profile were completely flat, would be shown as having all papers just above the borderline boundary. A very good candidate, achieving an A grade, would most probably have at least one paper in the exceptional band. In each paper a similar proportion of candidates fall in the exceptional and weak bands.

The profiled result slips attempt to achieve a useful compromise between the need to provide more information about performance in components, and a full-blown system of component-level grading. This latter option, as explained above, is not wholly appropriate for the construct of English language proficiency embodied in the UCLES EFL main suite exams. Like any compromise it is not without problems, one of which is that users may be tempted to over-interpret small and statistically insignificant differences between profiles. However, feedback from the trialling of the new-style result slips has generally been extremely positive.

Studies in Language Testing

The 10th volume in the Studies in Language Testing Series, *Issues in computer-adaptive testing of reading proficiency* by Micheline Chalhoub-Deville, is now available from bookshops. We have reproduced the series editor's notes by Michael Milanovic, Deputy Director of UCLES EFL.

The use of computers in language assessment has been a topic of great interest for some years and this volume makes an important contribution to thinking on computer adaptive testing (CAT) and reading comprehension. It considers the issues from a number of angles – reading research, design, development and measurement. The three main sections of the book are usefully reviewed by three discussants, Charles Alderson, Carol Chapelle and Bruno Zumbo who provide valuable insights through their comments, and the volume as a whole is ably edited by Micheline Chalhoub-Deville.

At Cambridge, much resource has gone into the development of both adaptive and linear computer-based tests. Work in this area started in the mid-nineties with a project to develop a CAT specifically for Manpower Europe, part of Manpower Inc. the world's largest employment services company. Linguaskill, as it is known, focuses on language for work purposes but is also notable for the fact that it is a multilingual system operating in English, French, German, Spanish and Dutch and reporting on the same measurement scale. Nine item types are used in Linguaskill, up to five of them focussing on reading. Two additional multilingual adaptive tests have also been developed in Cambridge – the computer based Business Language Testing System (CBBULATS) and CommuniCAT. The International English Language Testing System (IELTS) has also been computerised, though this test is linear rather than adaptive.

The development of all these tests have posed interesting practical and theoretical problems related to the way materials are presented, the interaction between test takers and the computer presentation of materials and how best to exploit the computer's power. The UCLES team has worked closely with the Multimedia Development Unit at Homerton College, Cambridge, which has significant expertise in educational software design.

Attention has also focused on using the computer to investigate the relationship between candidates' background characteristics, learning style, cognitive and metacognitive processes and test performance. This work

builds on that done by Jim Purpura, reported in Studies in Language Testing 8 and is intended to provide a resource for both learners and teachers. In addition, work continues in the area of self assessment and linguistic audit where the can-do system of performance descriptors, developed in nine European languages to date, is being computerised as part of a wider project to develop a multilingual performance-oriented descriptive framework of competence. The latter project will be reported in a later volume in this series.

Titles in the Studies in Language Testing Series are available from bookshops, or Cambridge University Press.

1 Lyle F Bachman, F Davidson, K Ryan, I-C Choi *An investigation in the comparability of two tests of English as a foreign language: The Cambridge – TOEFL comparability study*, Cambridge, 1995 (ISBN 0-521-48467-7)

2 Antony John Kunnan *Test taker characteristics and performance: A structural modelling approach*, Cambridge, 1995 (ISBN 0-521-48466-9)

3 Michael Milanovic, Nick Saville *Performance Testing, Cognition and Assessment: Selected papers from the 15th Language Testing Research Colloquium,* Cambridge and Arnhem, Cambridge, 1996 (ISBN 0-521-484465-0)

4 Caroline M Clapham *The development of IELTS: A study of the effect of background knowledge on reading comprehension*, Cambridge, 1996 (ISBN 0-521-56708-4)

5 Alison Green *Verbal protocol analysis in language testing research: A handbook*, Cambridge, 1998 (ISBN 0-521-58635-6)

6 *Multilingual glossary of language testing terms*, Cambridge, 1998 (ISBN 0-521-65877-2)

7 Alan Davies, A Brown, C Elder, K Hill, T Lumley, T McNamara *Language testing dictionary*, Cambridge, 1999 (ISBN 0-521-658764)

8 James E Purpura *Learner strategy use and performance on language tests,* Cambridge, 1999 (ISBN 0-521-658748)

9 Antony John Kunnan *Fairness and validation in language assessment*, Cambridge, 2000 (ISBN 0-521-658748)

10 Micheline Chalhoub-Deville *Issues in computer-adaptive testing of reading proficiency*, Cambridge, 2000, (ISBN 0-521-653800)

11 Catherine Elder (ed) *Experimenting with uncertainty* (ISBN 0-521-7725560) (in press)

12 Cyril Weir, Yang Huizhong, Jin Yan *An empirical investigation of the componentiality of L2 reading in English for academic purposes* Cambridge 2000 (ISBN 0-521-652995)

Forthcoming titles:

13 Kieran O'Loughlin *An investigatory study of the equivalence of direct and semi-direct speaking tests*

14 Anne Lazaraton *A qualitative approach to the validation of oral language tests*

Partial Competence Testing – the Certificates in English Language Skills (CELS)

The notion of partial competence in another language is now an important consideration in language learning around the world. For example, many learners may acquire comprehension skills (passive knowledge in listening and reading) without productive ability; alternatively learners may acquire skills in order to communicate orally without knowing how to read or write the language. Educators are realising that their approach to teaching (and therefore testing) needs to take account of this.

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.

CELS was developed from, and will replace, the Certificates in Communicative Skills in English (CCSE) and the Oxford EFL Reading and Writing Tests. Because of the similarities between CCSE and Oxford EFL, UCLES EFL decided to amalgamate the two examinations as part of the revision process. CELS now has the following features:

The three levels are linked to the underlying common scale for Cambridge EFL examinations:

CELS Preliminary – ALTE level 2

CELS Intermediate – ALTE level 3

CELS Higher – ALTE level 4

The writing and speaking modules use assessment criteria based on the UCLES Main Suite examinations. Research carried out during the revision project has shown that this allows the exams to be benchmarked to the criterion levels (to the UCLES/ALTE scale) and the reliability of the examinations has been enhanced (e.g. the speaking test includes two independent ratings).

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. The development of CELS will be the focus of a forthcoming volume in the Studies in Language Testing Series. Other UCLES projects – the CPE, BEC and IELTS Speaking Test revisions – will also be featured in the series.

European Year of Languages Conference

'European language testing in a global context' will provide an unprecedented opportunity for testing specialists, researchers and language teaching professionals from all over Europe and beyond to discuss the issues of language assessment on a global scale, covering a wide range of European languages. It will be one of the largest conferences on language testing ever to be held in Europe, and will cover three main strands:

European projects

the role of language testing as elaborated in the Common European Framework and in projects which have used this in practice; the European Language Portfolio project; projects for the development of European language tests; reports on projects funded by the European Union.

Language testing theory

reports on research projects; presentations related to developments in language testing process and delivery.

Use of language tests

issues of recognition by official bodies; supporting lifelong learning; increasing mobility; the role of IT in language testing.

codes of practise and ethics.

Speakers from throughout Europe, as well as Australia and the United States, will address a wide range of issues, from differing national perspectives and with reference to assessment in over 23 languages.

Presentations

Professor Charles Alderson, Lancaster University, UK
Current and future trends in language testing

Professor Lyle Bachman, UCLA, USA
The use of new technologies in language assessment

Dr Wolfgang Mackiewicz, Freie Universität Berlin, Germany
European language policy in the next decade

Dott. Ispettore Raffaele Sanzo,
Ministero della Pubblica Istruzione, Italy
Foreign languages within the frame of Italian educational reform

Mr Joe Shiels, Council of Europe

The role of the Council of Europe in European language policy: the portfolio for Young Learners

Dr John Trim, Council of Europe

The Common European Framework and its implications for language testing

The delegate fee for the conference is 200 euros, including lunch and refreshments. For more information on reserving a place at the conference, please contact:

Valérie Carré
ALTE Secretariat
1 Hills Road
Cambridge, CB1 2EU
United Kingdom

Phone: +44 1223 553260
Fax: +44 1223 553036
E-mail: carre.v@ucles.org.uk

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 553822
Fax: +44 1223 553068
e-mail: harding.a@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: