ResearchNotes

## Contents




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

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


## Editorial Notes

Welcome to issue 23 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The theme of this issue is technology, taken in its broadest sense. Technology is a key factor in maintaining our leading position in providing language assessment products and teaching awards. This issue provides an update on some major areas of work that we are involved in.

In the opening article Neil Jones gives a unique view of assessment systems, describing the complex system of language assessment in terms of its conceptual, human and technological facets that have evolved over the last decade at Cambridge ESOL. He describes the work of the many thousands of people worldwide who contribute to the running of our examinations and teaching awards as having a 'common purpose' which is to deliver language assessment of the highest possible quality. He outlines the contribution which each of the articles in this issue makes towards improving our provision of assessment through a diverse number of systems.

Helen Marshall then describes key aspects of our Item Banking System, the database of test items which are used to construct all of our examinations and tests. Helen describes how this system maintains quality assurance through a number of different stages that all test material goes through, drawing on current technology to do so.

Next Clare Mitchell Crow and Chris Hubbard describe how Cambridge ESOL is developing a web-based resource to support and communicate directly with all of our Professional Support Network (the worldwide community of examiners, item writers, presenters, inspectors and other external resources who provide professional support to Cambridge ESOL activity). The *ESOL Professional Support Network Extranet* will be of benefit to both the cadre of externals and Cambridge ESOL, allowing the provision of up-to-date information to, and feedback from, a large number of people and access to online communities for the participants.

Peter Falvey and Stuart Shaw continue a series of articles on the IELTS Writing Revision Project. They report the latest trial of new Writing assessment criteria and describe how well they are being interpreted and applied, followed by a discussion of the process approach to developing tests which examines how tests fulfil their intended purpose.

Staying with IELTS, we then provide performance data for IELTS test papers and candidates for 2004 followed by the announcement of the 2005 Masters Award winner and 2005 Funded Research Program recipients, together with the call for entries for the 2006 Masters Award. We follow this with a review of the 2004 Special Circumstances provision for candidates with special needs and an update of three key areas of research and development: computer-based testing (CBT), Asset Languages and ILEC, all of which are using technology in one form or another.

We end this issue with the latest conference reports, including Language Testing Forum hosted by Cambridge ESOL in November and some recent publications.

# Assessment systems: conceptual, human, technological

**NEIL JONES**, RESEARCH AND VALIDATION GROUP

## Introduction

Issue 23 of *Research Notes* covers the usual broad range of topics, providing an update on current developments within Cambridge ESOL. The thread I can use to draw them together in this introductory article is the notion of systems. A system is, to quote the *Compact Oxford Dictionary*, 'a set of things working together as a mechanism or interconnecting network'; or more particularly, in the context of computing, 'a group of related hardware units or programs or both,'; or more abstractly 'an organized scheme or method'. The papers in this edition describe systems of all such kinds – conceptual, human and technological – providing examples of how they interact and come together to implement the single, very complex system which is language assessment as practised by Cambridge ESOL.

Of course, administering tests to one and a half million candidates a year in 135 countries is a major logistical operation which requires the resources of a large organisation: over 200 staff employed directly in Cambridge, thousands more engaged in materials development and marking, not to mention the network of centres and the team of 15,000 conducting oral tests worldwide. Supporting this operation also evidently calls heavily on information technology (it is hard nowadays to imagine the days when the names and grades of every candidate were inscribed in red ink in large ledgers, though the ledgers are still preserved in the basement at the Cambridge-based head office, 1 Hills Road). But there is more to it than mere size or technical sophistication. Cambridge ESOL brings together people driven by a common purpose: to serve learners and teachers by delivering language assessment of the highest possible quality.

## Concern for pedagogy and positive impact

Since I joined the organisation in the early 1990s Cambridge ESOL has developed and changed beyond recognition. A theme that provides continuity has been the concern with language pedagogy and positive impact on learning, including the insistence on testing performance and the important role of human raters. Changes have included the development of pretesting and live response data capture, leading to greater reliability and consistency, and underpinning the use of measurement models to develop richer interpretative frameworks. Ways of testing have changed too: the first computer-adaptive tests date from 1993, beginning a strand of development the latest product of which is the online version of PET (see the update on CBT developments in this issue on page 20). The range of exams offered has expanded greatly, to cater for all age groups and different professional purposes. This development and diversification has led us to focus more on the validity of tests in terms of their fitness for particular purposes. This then is the complex system which I mentioned above: a wide range of language assessments, their operation and their quality supported by various technological means and by the application of organised human expertise.

## Conceptual Framework

At the heart of this system is the conceptual framework we use to think about test validity and the interpretation of test performance. This framework continues to change and develop. In the 1990s it incorporated theories of communicative language ability, and the four-part model of test usefulness (validity, reliability, impact, practicality proposed by Bachman (1990) and Bachman and Palmer (1996)). It can be seen to reflect the notions of evidence-centred test design (Mislevy 2003). It is now exploring Weir's (2005) socio-cognitive framework as a way of organising validity evidence within the test development process (see Shaw and Falvey's article on IELTS writing on page 7). A current issue is how validly to implement through assessment multilingual proficiency frameworks such as the Common European Framework of Reference or its home-grown British counterpart the Languages Ladder (see the *Asset Languages* update on page 20).

The papers in this issue all throw light on conceptual, human or technological aspects of the complex system that is Cambridge ESOL. Helen Marshall updates us on the *Cambridge ESOL item banking system*. This is the system at the heart of all test construction in Cambridge ESOL, and it implements the statistical measurement model which provides consistency in grading and in the interpretation of exam performance. Helen does not stress this aspect of the system, however, but rather shows how it lays down a workflow that supports quality assurance through the different stages of the test construction cycle. This is another demonstration of technology playing a supporting role within human systems.

In their presentation of the *ESOL Professional Support Network Extranet* Clare Mitchell Crow and Chris Hubbard describe the plan for a web-based approach to supporting approximately 20,000 external resources around the world. The first phase will concern support for the 15,000 Oral Examiners and Team Leaders, while the second will target item writers, written paper examiners, seminar presenters, chairs, Principal Examiners and centre inspectors. It provides a very interesting view of how technology in assessment can be used not to supplant human systems but to support them, providing a mechanism for professional development, promoting alternative working methods, creating a sense of community, and so on.

Peter Falvey and Stuart Shaw describe the completed project to revise the assessment criteria and scales for *IELTS Writing*. The focus again is on improving the working of human systems, this time by improving the performance of raters. Their paper succeeds in communicating the meticulous care with which this revision was undertaken and introduced.

Mike Gutteridge reviews *Cambridge ESOL's 2004 special circumstances provision* with respect to the Upper Main Suite of exams. Special circumstances cover the three areas of special arrangements:

- special arrangements for candidates with disabilities
- special consideration for candidates affected by adverse circumstances
- malpractice, i.e. cases of cheating.

As the article makes clear, this area involves making careful accommodation to individual circumstances. It is an area which is supported by technology in various aspects, and which will be so increasingly in future as computer-based testing develops and addresses accessibility issues.

Louise Maycock's update on *computer-based testing* gives a glimpse of the current wide range of technology-based developments in Cambridge ESOL. Karen Ashton's update on *Asset Languages* reminds us of the practical and conceptual challenges which this complex system must meet as it implements the DfES Languages Ladder. Critical here are standard-setting methods which can succeed in linking very different languages and groups of learners into the same functional proficiency framework. The focus is on developing learner-centred, can-do based approaches.

Standard-setting figures again, in relation to the use of IELTS for accrediting the language skills of health professionals, among the *reports from conferences*. Jay Banerjee and Lynda Taylor presented papers on this at LTRC at Ottawa in July 2005. Other conferences reported on include BAAL (September 2005 in Bristol), the LTF (Language Testing Forum), hosted by Cambridge ESOL (November 2005 in Cambridge), and the ALTE conference hosted under the auspices of the UK Presidency of the European Union (November 2005 in Cardiff). My own contribution to this section includes a presentation in Cardiff on recent benchmarking conferences in Sèvres and Munich, where international groups rated video samples for French and German on the CEFR. This presentation also relates to my theme rather well: a human system taking shape as raters strive to construct a shared understanding of a conceptual system (the CEFR levels), while technical or technological systems (the text of the scales, the format of the video samples) either aid or frustrate their efforts.

Cambridge ESOL continues to develop and refine its systems and procedures in the areas outlined above, in order to improve provision of language assessment to our stakeholders.

**References**


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

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

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

Mislevy, R (2003) On the Structure of Educational Assessments, CSE Technical Report 597, University of California, Los Angeles: Centre for the Study of Evaluation.

Weir, C (2005) *Language Testing and Validation: an evidence-based approach*, Oxford: Palgrave.


# The Cambridge ESOL Item Banking System


**HELEN MARSHALL**, ESOL PROJECTS OFFICE


In this article I will describe our current Item Banking System, updating somewhat the account found in *Research Notes* 1, considering the different parts of the Item Bank and what role they play in contributing to a quality assurance system.

## What is an item bank?

An item bank is, typically, a large collection of test items which have been classified and stored in a database so that they can be selected for use in tests. The items are all classified according to certain descriptive characteristics such as the topic of a text, the testing point for an item, etc., as well as statistical information about how difficult each item is. All of the item difficulties are located on a common scale of difficulty so that any combination of items can be put into a new test and the item difficulties combined to give a precise measure of the difficulty of that test. Items can be stored as discrete items or within tasks based around a stimulus, for example, multiple choice items based around a reading passage.

The security of test material is of paramount importance to Cambridge ESOL and our item banking system ensures that the material is as secure as possible. The database is encrypted, access rights are rigorously controlled and a number of secure passwords are required for Cambridge ESOL staff to access the database.

## Quality control

Many organisations now endeavour to practise Total Quality Management by which they mean that they adopt a comprehensive approach to achieving quality in every aspect of their work. For Cambridge ESOL, this starts with ensuring that we know all about the different kinds of people who take our examinations and exactly what it is they need and expect when they enter an examination. Not surprisingly, we have identified issues of fairness and the usefulness of our qualifications as key requirements of our examinations. Part of what fairness in language

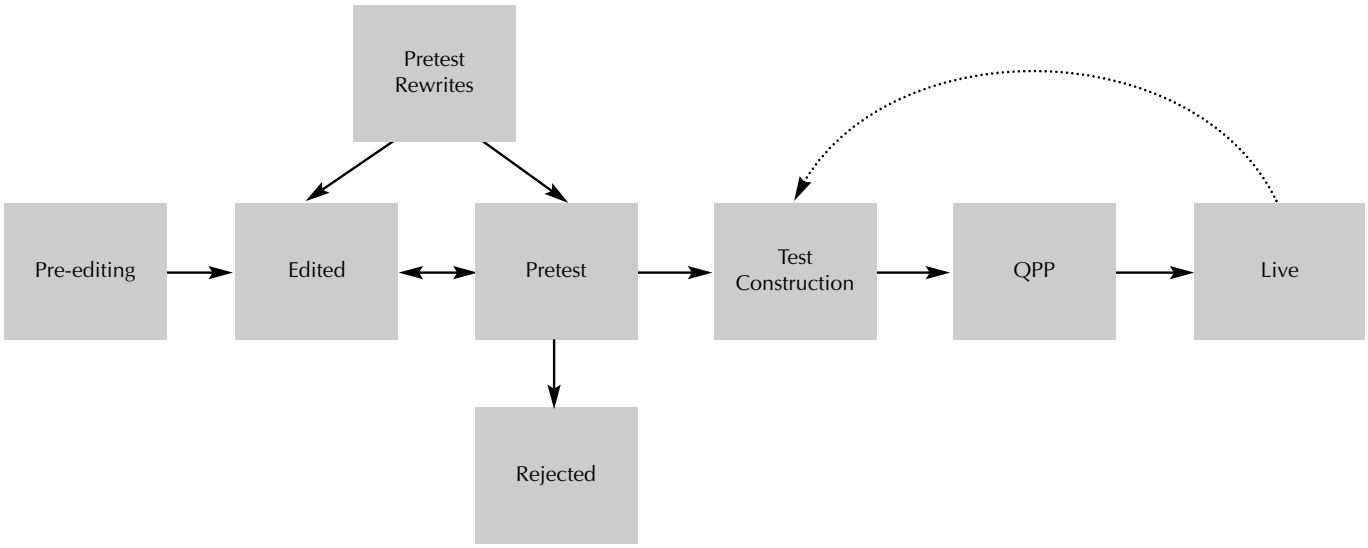


**Figure 1: Workflow through Cambridge ESOL's Item Banking System**

testing means is making sure that procedures for every stage in the testing process are well planned and carefully managed. This includes the way each test is produced and the way it is administered, marked and graded. Our approach to item banking addresses the way the test is produced by guaranteeing that all test material goes through a series of specific quality control checks before a completed test is constructed and administered.

## Bank system

The bank system used by each examination within the item banking system is effectively a workflow. This allows us to monitor material production to ensure that all the quality control checks made during the test construction process are completed to the required standard. Figure 1 shows the current architecture and workflow of the different banks within the Item Banking System.

The purpose of each bank is explained below.

*The Pre-editing bank*

Once material has been commissioned from Item Writers and received by Cambridge ESOL it goes through pre-editing and editing stages to ensure that it is at the correct level to be pretested. Currently, it is only at this stage that it is put into the item banking system and the material is word processed and stored in Microsoft Word. However, a new piece of software is currently being built by Cambridge Assessment, our parent organisation (see www.CambridgeAssessment.org for more information) that will allow Item Writers to write material on their PCs at home. Item Writers will use this software to accept commissions, write material and submit the material to Cambridge ESOL electronically. All material written in this way will be stored as XML (eXtensible Markup Language) which allows the material to be used in both paper and pencil and computer-based tests and to be re-purposed for any other use in the future.

The material will be downloaded from a web portal straight into the pre-editing bank where it will remain whilst it goes through the pre-editing and editing stages. Once it is ready for pretesting it will be moved to the Edited bank.

*The Edited bank*

From the material available in the Edited bank, the Chair of the paper decides how the material should be combined into pretests. The pretest booklets are created automatically by the system by combining all the required tasks and associated documents (front cover, examples, blank pages etc.) into a test document. The test document is then converted to PDF and a hard copy proof requested from the printers. Once the proofs have been printed and checked by the Subject Officer the tests are moved into the Pretest bank. The tests are then printed and sent out to centres worldwide.

*The Pretest bank*

The pretests sit in the Pretest bank awaiting the upload of item level statistics obtained from candidate responses. Both Classical item statistics (facility and discrimination) and item difficulties are uploaded. The item difficulties are derived using Rasch analysis, which relates the items to each other on the basis of common items (anchor items). The item difficulties are therefore anchored to a common scale thus making it possible to recombine tasks in the item bank and calculate how difficult a test will be for its intended candidature.

Once the statistics have been loaded the Pretest Review meeting can be held, which is where the material is reviewed in light of the statistical information and candidate/centre feedback. This stage is therefore the next quality control stage where the material is evaluated on the basis of whether the individual items discriminated between the stronger and weaker candidates and if the items were of the appropriate level of difficulty. It is at

this stage that the soft feedback is reviewed to ensure that there were no problems in terms of the suitability of the material. At the meeting material can be rejected if unsuitable, returned to the Edited bank to be rewritten and pretested again, or moved into the Test Construction bank.

*The Test Construction bank*

All the material in the Test Construction bank is at a known level of difficulty and can therefore be used to construct live tests. The Subject Officer and Chair of the item writer team meet and use the item banking system, on a networked PC, to select combinations of tasks that they may wish to include in a test. For any possible selection, the predicted overall difficulty of the test is immediately calculated by the system to aid the test construction process. This allows new versions to be constructed that are equivalent to previous versions and thus address a fundamental issue in testing; that of fairness. A report listing all the information held against the material describes the content of the test and allows the test constructors to determine whether the test adequately covers the required range of testing points. Once the tests have been agreed, and created, there is an Examination Ratification meeting where all of the papers for a particular administration are looked at together to consider the overall coverage of the whole examination. If no changes are required, the test is moved to the QPP bank where it begins the process of Question Paper Production (QPP).

*The QPP bank*

Once in the QPP bank, the paper production schedule for the test is started. The schedule automatically calculates the dates by which each event in the schedule needs to be completed if the test is to be ready on time. The first stage is the creation of the PDF version of the test followed by stages such as vetting, proofing and, finally, printing. An average QPP schedule has around 15 quality control stages which need to be signed off as complete in the item banking system. If a stage is not checked as complete by the required date, the item bank server automatically sends an email to staff responsible for the examination, alerting them. Once all the quality controls have been completed, the question paper is approved for printing. When the test papers have ben printed the test is moved to the Live bank.

*The Live bank*

The Live bank is effectively an archive of tests. Once a test has been taken, the item level statistics are uploaded and some tasks may be reused.

As described above, the Cambridge ESOL item bank is used to create tests, as well as store test items, and it has recently been developed to enable it to produce computer-based tests as well as paper and pencil tests.

## Item banking for computer-based testing

Cambridge ESOL has recently developed a computer-based testing (CBT) framework; the Connect Framework, that enables us to distribute, and administer, CB tests to centres around the world (see Seddon 2005 for further details). The framework allows centres to download CB tests, candidates take the tests and the centres then upload the information captured when the test was taken via a web portal back to Cambridge ESOL. Each candidate's performance can then be analysed, scored and graded as for paper and pencil tests.

In order to be able to render test items on screen for a CB test, the test items need to be stored as XML along with any necessary graphics and sound files. The Cambridge ESOL item banking system now has the ability to tag test items with the appropriate XML needed to be able to produce an electronic 'test bundle' that will enable the test items to be rendered on screen. In order to be able to proof the test on screen, so we can see what the candidates will see, there is in-built functionality that allows Subject Officers to view the CB task, and CB test, on screen.

## Conclusion

It would not be possible to conclude this brief account of the Cambridge ESOL test production methodology without saying something about how this process addresses issues of validity and reliability. Validation is often described as the process of building an argument to support the inferences that are made from test scores. For Cambridge ESOL examinations, that process is greatly assisted by the systematic review of new items throughout the stages identified above. Similarly, the pretesting stage identifies items which may be performing poorly for some reason. By removing these items, the remaining material is of a better quality and will measure more reliably. In addition to the extensive validation work carried out post hoc – after the test administration – using live data, the effort that goes into producing the tests can clearly be seen as contributing to the overall validity, reliability and, above all, the quality of Cambridge ESOL examinations.

**Reference**

Seddon, P (2005) An overview of computer-based testing, *Research Notes* 22, 8–9.

# ESOL Professional Support Network Extranet

**CLARE MITCHELL CROW**, ESOL PROJECTS OFFICE
**CHRIS HUBBARD**, PERFORMANCE TESTING UNIT

## Introduction

Running a large and complex range of public examinations requires a considerable amount of professional support around the world, and external resources are fundamental to the operation of Cambridge ESOL's examinations. The most obvious examples are examiners for speaking tests and writing papers. The Professional Support Network, in this context, is defined as consultants engaged on an individual basis to carry out a defined role or roles. Cambridge ESOL currently supports a variety of professional development activities, however, there is currently no active web-based resource offering to support and communicate directly with these external consultants.

Developing a web-based tool that offers professional support for the professional support network will meet the twin objectives of: allowing Cambridge ESOL to establish and maintain a direct relationship with those external resources who support the examination procedures, and providing the opportunity to access up-to-date support materials that are key to their roles.

## Why a Professional Support Network Extranet?

Developing a Professional Support Network (PSN) extranet will not only allow Cambridge ESOL to establish a direct relationship with those external resources that support Cambridge ESOL examinations, it will also:

- Provide a mechanism for ongoing professional development and support
- Promote alternative working methods
- Allow our PSN to establish direct links with their peers, and create a sense of community
- Improve efficiency and decrease cost of material flow by providing downloadable materials
- Promote Cambridge ESOL, and offer information on topics relevant to the PSN.

The Professional Support Network Extranet development will offer support via the Internet. With this mode of support, there will be enhancements to existing materials that are used by these professionals, including a dedicated website, discussion forums and interactive training and development tools.

## Who will be supported?

The Cambridge ESOL PSN comprises approximately 20,000 people around the word. The launch of the site will be managed in two main stages, with Phase 1 of the site delivering support materials for the Team Leader system and Oral Examiners, who number approximately 15,000.

Delivery of Phase 2 of the Professional Support Network will eventually ensure that the following roles have online access to support material: Item Writers, Written Paper Examiners, Seminar Presenters, Chairs, Principal Examiners and Centre Inspectors.

### Phase 1 of the PSN development

The aim of the PSN development is to support both the global cadre of examiners and the implementation of quality assurance procedures by providing more direct support in a more flexible and user-friendly manner. Oral Examiners are responsible for managing Speaking test procedures and assessing live candidate performances. The international standardisation of examiner conduct and assessment is managed by the Team Leader system. The system is based on a well defined and tested set of procedures called Recruitment, Induction, Training, Co-ordination, Monitoring and Evaluation (RITCME). The intention of introducing the PSN is to give Centres, Team Leaders and Oral Examiners more options for how they complete the stages of RITCME.

Examples of the ways in which the development of the PSN aims to achieve this goal are described below:

*Providing induction materials*

The present Induction stage of RITCME uses a self-access pack consisting of a video and worksheets. Presently these are sent to new examiners and have to be returned to the centre after use. Providing these materials online will allow a number of new examiners to have access to the materials concurrently and will negate the need to send out and return materials. These resources would remain constantly available to examiners so that anybody wishing to refresh their knowledge could access them as needed.

*Supporting the streamlining of Training and Co-ordination sessions*

By providing online materials and updating meeting guidelines accordingly, these stages of RITCME will have a reduced face-to-face meeting load. Examiners using the PSN Extranet will still cover the same amount of material in a similar timescale but will have more flexibility about when they do so. By completing online exercises and assessments, examiners will be able to chose when they do the work and at what speed. The focus of shorter face-to-face meetings will be on the stages of development that need practical application in a group environment.

*Offering on-going support*

A great advantage of providing a number of materials in an online environment is the ability to give examiners access to 'development' materials and extra assessment practice. Examples of this will be video clips and activities that focus on aspects of examiner procedure and give examples of best practice, or

assessing past standardisation clips. Examiners could be directed to these by their Team Leader as required, or access them individually.

*Building links*

The system will not only allow direct and ongoing contact between Team Leaders and their teams of examiners but will also allow Cambridge ESOL to update members of the Team Leader system on relevant exam developments. It will also be much easier to canvass examiner opinion on appropriate questions and to foster more of a feeling of inclusion amongst a global and diverse group of professionals.

## Impact of the PSN Extranet

The introduction of this system will see major changes in the way that Cambridge ESOL's PSN is trained and supported. Rather than completely replacing the current systems of supporting external resources the function of the PSN will be to enhance and improve the support that Cambridge ESOL currently provides. Face-to-face training and support, and the provision of Handbooks, Guideline manuals and videos/DVDs will all still have a role to play in an overall support package. It is hoped that the benefits of this type of provision will enrich the professional development experience of our PSN.

The changes will occur gradually, allowing for the stakeholders to prepare for the changes in both technology and process. In Phase 1, the changes to how RITCME is delivered offer a range of opportunities and constraints. These include:

*Opportunities*

- **Access** – Offering the development over the internet allows for 24/7 delivery of material, and instant updates to new material
- **Flexibility** – The system will be flexible to ensure participants can learn at their own pace, and participate in the development activities when it suits them
- **Availability** – Participants will be able to learn at a time and a place that is convenient to them
- **Autonomy** – Offering professional development in this mode offers participants a certain level of freedom, to take responsibility for their own development
- **Individualised Learning** – This method of delivery allows individuals to tailor the learning experience to themselves and interact with their Team Leader
- **Community Building** – Allowing members of the PSN to establish and foster links in communities that share ideas and increasing the opportunity for interaction between both external resources and Cambridge ESOL.

*Constraints*

- **Access** – Initially, it is recognised that some participants will have insufficient access to the internet to enable participation in advanced features of the system such as watching training videos
- **Isolation** – Although the system will encourage participation and interaction via forums and email, it will also cause participants to meet face-to-face for shorter periods than under the current regime.

## Summary

The introduction of the PSN website provides Cambridge ESOL with an exciting opportunity to engage with our external resources in a way that the organisation has not previously explored. The challenge is to ensure the development and delivery of an engaging, professional development environment.

Phase 1 of the Professional Support Network will be available in 2006 to selected external resources and Phase 2 will follow in 2007–08. Further information about this development will be included in future issues of *Research Notes*.

# IELTS Writing: revising assessment criteria and scales (Phase 5)

**PETER FALVEY**, CAMBRIDGE ESOL CONSULTANT
**STUART SHAW**, RESEARCH AND VALIDATION GROUP

## Background

The new IELTS Writing assessment criteria and scales were used operationally for the first time in January 2005. A trial was undertaken in April with a group of highly experienced IELTS examiners from Anglia Polytechnic University (UK) in order to ascertain how well the revised assessment criteria and band level descriptors are functioning. The trial entailed a preliminary discussion of examiners' initial thoughts prior to script marking; verbal protocols, where an examiner was asked to think aloud while actually marking, and a focus group discussion after marking was over, mainly covering reactions, attitudes and behaviour of examiners to the new scale. It was hoped that the trial would offer greater understanding of how well the revised rating scale is interpreted and applied by examiners which will provide further validation evidence to support its introduction.

Cambridge ESOL has designed and successfully implemented a model of test development which comprises, as essential to its functioning, research and validation dimensions. The model

adopts a process-based approach to development and revisions. The approach is procedural in that it specifies several phases which must be adhered to and followed through in order to develop and validate tests or examinations. The underlying approach – being both cyclical and iterative – is based on the need to establish the 'utility' of a test in fulfilling its intended purpose in a useful way. More traditionally, this is thought of as being the overriding principle of validity as defined as '*fitness for purpose*'. Four essential qualities of test or examination usefulness, collectively known by the acronym VRIP (*Validity, Reliability, Impact* and *Practicality*), have been identified as aspects of a test that need to be addressed in establishing fitness for purpose. The four elements of VRIP are considered by Cambridge ESOL to contain the essential constructs and elements of a 'good' test either in development or in revision. The underlying Cambridge ESOL model is realised through the five phases of test development and revision adopted by the IELTS Writing Assessment Revision Project which have been reported in *Research Notes*:

- Phase 1: Initial Planning and Consultation – *Research Notes 9*
- Phase 2: Development – *Research Notes 10 and 15*
- Phase 3: Validation – *Research Notes 16*
- Phase 4: Implementation – *Research Notes 18*
- Phase 5: Operation.

The constructs of *Validity* and *Reliability* underpin Cambridge ESOL's development and revision model. The other two elements in the test model, *Impact* and *Practicality* are also important features of a test. *Impact* refers to the effect of the new or revised test on its stakeholders and the feedback that they provide to test developers. In the case of the revision project, the major stakeholders were: the test developers, the three IELTS partners, and the clients. The clients consist of test-takers and the institutions to which the test-takers would apply for academic and/or professional and vocational programmes. Other stakeholders, to a certain degree, would be the raters for the Writing test, text-book writers, publishers and those who prepare potential test-takers for IELTS through programmes of learning. Throughout the revision project a wide variety of stakeholders have been included in all the phases of the revision. The approbation with which examiners have welcomed both the training and examining experience was especially apparent during the *Implementation Phase*. *Practicality* is an important element in test revision because this element focuses on whether the objectives of the test can be met without major constraints, inconveniences and other logistical problems.

## IELTS Writing Assessment Revision Project: the story so far

The decision to follow the five phases of the project meant that a formal, thorough series of processes were followed throughout the revision process. Although there was a certain amount of overlapping in some of the phases, notably the *Validation* and *Implementation* phases, the phases were, on the whole separate and distinct. This allowed the IELTS partners, the IELTS Steering Committee, the Writing Assessment Revision Working Group, test developers and examiners to proceed through the phases in a deliberate and coherent manner, retaining the integrity of the project objectives.

A number of issues that were well addressed throughout the project included the substantive consultation with stakeholders and the iterative procedures that occurred extensively, to amend, change and 'tweak' assessment criteria, descriptive bands, scales, and rubric.

In addition to the iterative process and the relatively large-scale validation exercises, a key feature of the project has been the willingness of the project developers to seek to inform themselves and push forward the test revision process at any stage of the revision development process by:

- a combination of further trial exercises when one trial was considered insufficient
- the commissioning of research projects when it was clear that some research should be carried out to inform the Working Group (e.g. the study on changes to the rubric)
- the compilation and analysis of literature reviews.

What this means in practice is that the project can never be described as a one-shot effort. The revision of a high-stakes examination should never be approached by means of a monolithic exercise without the opportunity to go back, to seek further insights and to be willing to adapt during the process by revising, amending or rejecting previous decisions in the light of further data analysis, the analysis of which sometimes requires rethinking.

In discussing the *Validation Phase*, the number of iterations and the large number of assessments that were carried out by raters plus the analysis of the quantitative and qualitative data that were produced by the large-scale exercise revealed a serious approach to issues of validity and reliability. The best performance test can fall at the twin hurdles of validity and reliability. It is clear that issues of validation in a performance test are vital but it is also clear that no matter how good the validity of the test, unreliable assessment will destroy all the good work that has gone before. The analysis of the data collected as part of Phase 3 revealed that the reliability had been addressed in a satisfactory manner.

As the revision project progressed, the issue of the training and standardising of writing raters became vitally important. Phase 4, the *Implementation Phase*, needed to show that raters could be trained and standardised sufficiently well to make consistent judgements on rating exercises. The amount of time and effort put into developing a validation resource for Phase 3, that later, after a number of iterations and amendments, was adapted into a comprehensive training and standardisation programme, showed how important the Working Group considered the issue. The IELTS Writing Test (and any other high-stakes performance test) will be considered invalid if the inconsistent rating of scripts occurs. The decision to retrain and re-standardise all IELTS Writing examiners had large implications for the practicality element of the examination in terms of time and expense. Nevertheless, the Working Group made the decision to produce a thorough training package that was first tried out on senior examiners, amended in the light of the feedback that was received, and trialled again with experienced raters who, having satisfactorily completed the

training package, were able to move out and use cascade training and standardisation for all existing and potential raters.

The existing requirement for raters to be certified was enhanced by the decision to train and re-certify all IELTS Writing test raters thus ensuring high *impact* and the enhancement of the validity and reliability of the examination.

## Trial methodology

One clear and immediate area for the focus of future studies involves an investigation into the efficacy of Phase 5 – the *Operational Phase*. Phase 5 is the culmination of the development work and training and standardisation of raters that has evolved since 2001. A trial took place in April 2005 with a group of IELTS raters to ascertain how well the revised assessment criteria and band level descriptors have been operating.

The trial addressed the following questions:

- What are the attitudes and reactions of examiners towards the new assessment scale?
- How well is the revised rating scale operating?
- Do examiners interpret and apply the new scale in a consistent manner?
- What are the issues raised by the new assessment approach?

Evidence collected from semi-structured, focus group discussions and individual verbal protocols make it possible to demonstrate that the revised rating scale is functioning in a valid and reliable manner. The study involved asking a small group of four experienced examiners to articulate their thoughts about the revised scale through a semi-structured, facilitated discussion in order to ascertain their views after operationally using the rating scale over a 3–4 month period. Previous validation studies (Shaw 2001, Raikes et al. 2004) have employed introspective techniques, particularly concurrent 'think-aloud' protocols to explore examiner cognitive processes. It is believed that individual 'think-aloud' protocols offer rich insights into how candidates and examiners make judgements and will contribute towards informing the operation and efficacy of the revised rating scale.

All examiner responses elicited during open discussion and throughout the verbal protocols were audio recorded. Audio recording of vocalised examiner thoughts as they individually rated tasks provided immediate and explicit explanations of:

- What do markers do as they are marking?
- How are markers' thought processes structured during the marking process?
- What particular information do markers heed when judging candidates' answers?

Profiles for each of the four trial examiners are presented in Table 1.

### Focus group discussion

The focus group discussion covered areas relating to initial script management, approach to assessment, use of band descriptors, paragraphing, old and new scale comparability, formulaic language, training, guidelines for word counts, and script legibility. Examples of questions guiding the discussion included:

**Table 1: Examiner background information**

| | E1 | E2 | E3 | E4 |
|---|---|---|---|---|
| **Experience as an examiner?** | **FCE, CAE, CPE, IELTS** | **PET/FCE/IELTS/ Skills for Life – Written** | **FCE, CAE, CPE – Oral and Written** | **IELTS Main Suite** |
| Years as an IELTS examiner | 15 | 2.5 | 7 | 7 |
| Years as an EFL/EAP teacher | 20+ | 10 | 15 | 10 |

- Do you rate Task 1 responses before you rate Task 2?
- Do you prefer to assess an entire candidate before moving on to the next?
- Do you tend to read responses more than once or not?
- Do you tend to concentrate on one assessment criterion or on all four?
- Are the new band descriptors clearly worded?
- Are you satisfied with the accuracy of your final award?

Some of the questions were based on observations, reports and concerns articulated by examiners participating in the global survey administered during Phase 1 of the project in 2001.

### Examiner questionnaire

Retrospective data was also captured by an examiner questionnaire. The small-scale survey of examiners aimed to deduce how they felt about the revised writing assessment procedures; how, as individuals, they rate *Academic* and *General Training* written responses to Tasks 1 and 2; how they felt about revised assessment criteria and band descriptors and their general impressions and opinions of the assessment process and training.

The questionnaire was designed to be completed in 30–40 minutes and questions were phrased such that they were appropriate to the revised task criteria for both *Academic* and *General Training* responses. The questionnaire consisted of five sections:

1) Examiner Background Information
2) Using the IELTS Revised Assessment Criteria and Band Descriptors – General Rating Issues
3) Using the IELTS Revised Assessment Criteria and Band Descriptors – Task 1
4) Using the IELTS Revised Assessment Criteria and Band Descriptors – Task 2
5) Other Issues.

The questionnaire – refined in the light of trial findings – has been circulated globally to thirty major IELTS test centres during summer 2005.

### Verbal Protocol Analyses

Verbal Protocol Analysis (VPA) is a methodology based on the assertion that an individual's verbalisations may be perceived as an accurate account of information that is (or has been) attended to as

a particular task is (or has been) undertaken. In other words, what examiners say is what they do, and that the type, direction and number of comments bears some relationship to the ratings awarded. Usually, the individual is asked either to *'think aloud'* or *'talk aloud'* as the task is carried out. Sometimes the individual is asked to verbalise concurrently (as the task is being carried out) or retrospectively (after the task has been carried out).

VPA differs from other techniques (discourse analysis or interviewing) – which focus primarily on linguistic content and structure and the information of what is said – in the sense that VPA offers the possibility of making inferences about the cognitive processes that produce the verbalisation. VPAs might contain the utterances made as the individual carries out a single task, or a series of tasks. A set of protocols might be gathered from different individuals completing similar tasks, or from the same individual completing different tasks. The set of protocols gathered constitute a body of qualitative data (Green 1998:1).

Verbal protocols may be gathered in slightly different ways and under varying circumstances, depending on the type of research question that is to be addressed. The generic verbal report procedure comprises *Form of Report, Temporal* and *Procedural* variations. The procedure adopted for this study constituted: *Form of Report – 'Talk aloud'* (in which individuals verbalise a task as it is carried out); *Temporal – Concurrent* (concurrent or simultaneous verbalisations which are generated at the same time as the individual is working on the task); and *Procedural Non-mediated* (where the individual is asked to talk aloud and is prompted only when he/she pauses for a period of time).

Two *Academic* Task 1 tasks and two *Academic* Task 2 tasks were used for the trial. *General Training* tasks were not used. The tasks differed in terms of the amount of information the candidates were required to process to complete the task and the degree of processing demanded of the examiners who were required to rate the responses to those tasks. In Task 1, examiners were either asked to look at a map showing the town of Garlsdon (Task 1a) or a multi-line graph (dates on x axis) showing information about the performance of a train company in October and November 2002 (Task 1b). Given that direct questions are fairly untypical Task 2 question types, it was decided to include the more conventional task for the trial i.e. '*to what extent …*' questions (Task 2c and 2d).

Commenting on a particular candidate response, examiners were invited to consider the following issues during rating:

- describe their individual method of assessment
- explain how they applied and engaged with the rating scale
- state which criteria they focused on first
- state which criteria they paid most attention to
- describe how they reacted to the response in the light of the revised rating scale
- comment on those features of a candidate response they believed differentiated between strong and weak Task 1/Task 2 performances
- identify the strengths and potential problems with the revised scale.

Verbal protocol analyses revealed individual assessment approaches for each of the four examiners (E1–E4), described in Figure 1.

Milanovic and Saville (1994:98–100) have investigated marking strategies using verbal protocols in Cambridge examinations. Their data revealed four discernible approaches to composition marking: *Principled* and *Pragmatic Two-Scan Approaches, Read Through Approach* and *Provisional Mark Approach.*

Markers adopting the *principled two-scan/read approach* scan or read the script twice before deciding on a final mark. The second reading is 'principled', being undertaken indiscriminately with all scripts, hence the term 'principled two-scan/read'. Markers adopting the *pragmatic two-scan/read approach* to the process of marking also read the scripts twice before assigning a mark to the script. What distinguishes this marking approach from the *principled two-scan/read approach* is the motivation behind the second reading of the composition. The *pragmatic two-scan/read* occurs only when the marker encounters difficulties in the script or in the marking environment and has to re-read to determine a mark. That is to say, markers only have recourse to this approach in the event of the failure of another method to generate a confident mark. *Read through* is the least sophisticated of the marking approaches and consists of reading a script through once to pick up its good and bad points. The *provisional mark approach* is also characterised by a single reading of the script, but with a break in the marking flow, usually imposed towards the start of a candidate's effort, which prompts an initial assessment of its merits before reading is resumed to discover whether the rest of the answer confirms or denies that assessment.

These approaches provide a means for understanding the behaviour of the trial examiners. IELTS examiners are currently required to analyse the task first. Clearly, it is important to rate all responses to a task before moving on to the second task. The four trial examiners appeared to analyse the task requirements before attempting to rate their responses. E1, E3 and E4 revisited the task throughout marking. Altogether more circumspect, E2 made every effort to understand the task thoroughly before attempting to rate the response. Previous surveys (Shaw 2002) have indicated that examiners may read tasks many times and make note of features students are expected to include in their response.

Encouragingly, it is evident from the protocols that examiners are generally adhering to the revised method of assessment (described in Bridges and Shaw, 2004). That is, examiners tend to begin with *Task Achievement* – Task 1 and *Task Response* – Task 2 (noting length requirements) before moving on to the other three criteria. For each criterion, each examiner begins with the over-arching statement that most closely matches the global features of the script. Examiners then read through the detailed features of performance at that band and attempt to match the features of the script – checking that all the positive features of that band are evident in the script. In order to 'fine-tune' their ratings, examiners seek to confirm or deny their awards by giving consideration to the descriptors both below and above the band. This enables them to ensure that there are no penalties/ceiling that are relevant and that the rating is ultimately accurate.

E1 adopted a *principled two-scan/read approach* to marking. E2 appears to combine marking approaches. She read the response at least twice before deciding on a mark for *Coherence and*


**Figure 1: Individual assessment approaches to rating writing performances**

*Cohesion* (*principled two-scan/read approach*) commenting:

> *'It takes longer to rate a script. Perhaps 3 readings. The first for a general overview, the second to choose bands in terms of overarching statements and the third to fine tune those bands in terms of detailed descriptions.'*

Interestingly, however, on her first reading of the response she formed a fairly confident evaluation of *Lexical Resource* and *Grammatical Range and Accuracy*. This approach conforms to the *read through approach*. E2 appears to adopt different marking approaches for different criteria.

During the rating of Task 1b, protocols revealed that E3's initial impression of the response was slightly modified on second reading (*pragmatic two-scan/read approach*). In this case an early impression was altered with a subsequent reading. E4's rating of Task 2d was characterised, at least in part, by a single read followed by selective reading. Protocols revealed long periods of silence in

which the examiner broke away from the response in order to scrutinise the band descriptors (the *provisional mark approach*).

## Trial Findings

Throughout the trial analysis, findings were related to observations, reports and concerns articulated by examiners participating in the global survey administered during the *Consultation, Initial Planning and Design* (Phase 1) of the project in 2001. Comparisons were also made with findings garnered from the main validation trial undertaken in 2003 when the revised scale was used by senior examiners for the first time.

Examiners find the revised rating scale a considerable improvement on the old scale and welcome the greater clarity and additional explanatory text in the new descriptors. Examiners also believe the revised rating scale provides a more comprehensive description of the key features of writing at each band level. The separation of *Lexical Resource* and *Grammatical Range and Accuracy* is perceived to be extremely valuable: a belief widely held by examiners throughout the *Implementation Phase*.

The verbal protocols indicated that raters are processing several assessment criteria simultaneously and that all four assessment criteria are uppermost in the minds of raters when evaluating either Task 1 or Task 2. Protocols also reveal that raters tend to:

- analyse the task requirements before rating responses
- revisit the task throughout their marking
- base their assessments on a detailed study of the features of the candidate's response in relation to the task
- employ different marking approaches for different criteria
- adopt one of two marking approaches: a *'principled two-scan-read'* and a *'pragmatic two-scan/read'*.

Encouragingly, the revised, prescribed method of assessment is being universally adopted. The revised approach is depicted as a flow diagram in the *Instructions to IELTS Examiners* booklet and is now a prominent feature of examiner training. Examiners understand the revised criteria for Task 1 and Task 2 and have acquired confidence in their ratings after using the scale operationally for three months. There is a general satisfaction with the accuracy of final awards.

The subscales seem to work well for each of the two writing domains. Additionally, the revised *Task Achievement/Task Response* criterion is effective for rating Task 1 and Task 2 across the differing domains. The use of emboldened 'ceiling statements' in the band level descriptors in relation to dealing with problem scripts is clearly effective. The penalty system for underlength scripts appears to be working well and examiners regard the guidelines for what constitutes a word and a 'word count' a helpful aid in assessing IELTS scripts.

The new scale is not without criticism, however. Examiners have suggested that the current and revised scales are not of the same standard as the new scale appears marginally more harsh than the old one. Examiners also regard the criteria for paragraphing as being somewhat strict believing that too much weight is given to paragraphing. It was thought that candidates who attend IELTS preparation classes and who are taught paragraphing have some advantage over those not preparing for the examination through traditionally-taught courses. Verbal protocols and examiner observation reveal that it is perfectly conceivable for candidates to produce a highly competent Task 1 in the absence of paragraphing yet the current descriptors would deny a Band 8 or 9. Additionally, the actual process of rating a script is marginally longer now than when rating under the old scheme.

The examiners who took part in the trial wholeheartedly commended the new training procedures and accompanying materials. However, they were less complimentary about Certification which is considered to be very stressful and its inclusion in training is perceived to be too heavy a burden. Trial examiners advocate that Certification should be experienced as a separate event. Moreover, training has occasionally failed to communicate effectively the principle of 'best fit' when matching performances against band descriptors.

## Conclusion

The preliminary *Operation Phase* trial has, on the whole, revealed that the revised IELTS Writing Scales, Bands and Descriptors are working well and are an improvement on the previous set.

This article, in conjunction with a series of IELTS revision articles published in *Research Notes*, has endeavoured to describe and discuss the processes and outcomes of the IELTS Writing Assessment Revision Project. We have revealed the considerations that were involved in the planning and development of the project, the studies that were undertaken in order to ensure the validity of the project and the reliability of the application of the revision changes, the ways in which the revisions were implemented and the extensive training that took place involving the training and re-training of all writing examiners.

A high-stakes examination will always move forward as the process of operations reveals problems and issues that have been missed during the formal development process.

Other studies undertaken throughout the *Operation Phase* of the project, including a global survey of raters and a study of rater responses to the re-certification process, will be reported on as they are undertaken throughout 2005 and 2006.

**References and further reading**


Bridges, G and Shaw, S D (2004) IELTS Writing: revising assessment criteria and scales (Phase 4), *Research Notes* 18, 8–12.

Green, A (1998) *Verbal Protocol Analysis in language testing research. A Handbook*. Studies in Language Testing Vol 5, Cambridge/UCLES: Cambridge University Press.

Milanovic, M and Saville, N (1994) An Investigation of Marking Strategies using Verbal Protocols, internal report, UCLES.

Raikes, N, Greatorex, J and Shaw, S D (2004) A Report on the March 2004 OCR On-Screen marking Trial: The examiners' experience, internal report, UCLES.

Shaw, S D (2001) The effect of standardisation training on rater judgement and inter-rater reliability for the revised CPE Writing paper 2, internal report no. 290, UCLES.

—(2002) IELTS Writing Assessment: Towards Revised Assessment Criteria and Band Descriptors. A Quantitative and Qualitative Analysis of IELTS Writing Examiner Feedback, internal report, Cambridge ESOL.

# IELTS test performance data 2004

Each year, multiple versions of each of the six IELTS Modules (Listening, Academic Reading, General Training Reading, Academic Writing, General Training Writing, and Speaking) are released for use by centres testing IELTS candidates. Reliability estimates for the objectively and subjectively scored Modules used in 2004 are reported here, together with candidate performance on both Academic and General Training Modules.

## Reliability of objectively-scored Modules (Reading and Listening)

The reliability of listening and reading tests is reported using Cronbach's alpha, a reliability estimate which measures the internal consistency of the 40-item test. The following Listening and Reading material released in September 2004 had sufficient candidate responses to estimate and report meaningful reliability values as shown in Table 1.

The figures reported for Listening and Reading Modules indicate the expected levels of reliability for tests containing 40 items. Values for the Listening tend to be slightly higher than those for the Reading components; both Academic and General Training candidates take the same Listening Module so the test population represents a broader range of ability.

On the basis of these reliability figures, an estimate of the standard error of measurement (SEM) may be calculated for these Modules[1]. The mean band scores, standard deviation and standard error of measurement for Listening and Reading Modules are shown in Table 2.

The SEM should be interpreted in terms of the final band scores reported for Listening and Reading Modules (which are reported in half-bands).

## Reliability of subjectively-scored Modules (Writing and Speaking)

The reliability of the Writing and Speaking Modules cannot be reported in the same manner as for Reading or Listening because they are not item-based; candidates' writing and speaking performances are rated by trained and standardised examiners according to detailed descriptive criteria and rating scales. The assessment criteria used for rating Writing and Speaking performance are described in the IELTS Handbook (and see Falvey and Shaw's article in this issue). The rating scale band descriptors are examiner-oriented scales and are not in the public domain; however, benchmarked example writing performances and CD-based speaking performances at different levels can be found, along with examiner comments, in the IELTS Specimen Materials[2]. User-oriented scales describing levels of Writing and Speaking performance are available.

1. See Analysis of test data section of the IELTS website for formulae www.ielts.org
2. See Downloads section of the IELTS website for handbook and specimen materials order form.

**Table 1: Reliability for Listening and Reading Modules**

| Module | Version | Alpha |
|---|---|---|
| **Listening** | 199 | 0.89 |
| | 200 | 0.90 |
| | 201 | 0.88 |
| | 202 | 0.90 |
| | 203 | 0.89 |
| | 204 | 0.90 |
| | 205 | 0.90 |
| | 206 | 0.91 |
| | 207 | 0.87 |
| | 208 | 0.87 |
| | 209 | 0.88 |
| | 211 | 0.87 |
| | 212 | 0.83 |
| | 213 | 0.90 |
| | 214 | 0.89 |
| *Average alpha across versions* | | *0.89* |
| **Academic Reading** | 199 | 0.84 |
| | 200 | 0.86 |
| | 201 | 0.90 |
| | 202 | 0.85 |
| | 203 | 0.88 |
| | 204 | 0.88 |
| | 205 | 0.88 |
| | 206 | 0.83 |
| | 207 | 0.86 |
| | 208 | 0.84 |
| | 209 | 0.90 |
| | 211 | 0.90 |
| | 212 | 0.87 |
| | 213 | 0.90 |
| | 214 | 0.84 |
| *Average alpha across versions* | | *0.86* |
| **General Training Reading** | 199 | 0.89 |
| | 200 | 0.90 |
| | 201 | 0.88 |
| | 202 | 0.87 |
| | 203 | 0.87 |
| | 204 | 0.88 |
| | 205 | 0.86 |
| | 206 | 0.90 |
| | 207 | 0.89 |
| | 208 | 0.90 |
| | 209 | 0.86 |
| | 211 | 0.83 |
| | 212 | 0.86 |
| | 213 | 0.89 |
| | 214 | 0.87 |
| *Average alpha across versions* | | *0.87* |

**Table 2 Mean, standard deviation and standard error of measurement for Listening and Reading Modules**

| | Mean | SD | SEM |
|---|---|---|---|
| Listening | 6.01 | 1.25 | 0.37 |
| Academic Reading | 5.87 | 1.04 | 0.39 |
| General Training Reading | 5.80 | 1.24 | 0.46 |

Reliability of rating is assured through the face-to-face training and certification of examiners and all examiners must undergo a retraining and recertification process every two years. Continuous monitoring of the reliability of IELTS Writing and Speaking assessment is achieved through a sample monitoring process. Selected centres worldwide are required to provide a representative sample of examiners' marked tapes and scripts such that all examiners working at a centre over a given period are represented. Tapes and scripts are then second-marked by a team of IELTS Principal Examiners and Assistant Principal Examiners. Principal Examiners monitor for quality of both test conduct and rating, and feedback is returned to each test centre. Analysis of the paired examiner–Principal Examiner ratings from the sample monitoring data collected and analysed for 2004 produced an average correlation of 0.89 for the Writing Module and 0.91 for the Speaking Module.

It is customary to use inter-rater correlations in calculating the reliability of subjectively marked tests such as the Speaking and Writing components of IELTS. Where a single rater is employed, the Spearman-Brown formula given on the IELTS website can be used to generalise to the case of a single rater from the correlation found between two ratings. This formula gives a reliability of 0.80 for the Writing and 0.82 for the Speaking Modules. Table 3 shows the mean band scores, standard deviation and standard error of measurement for Writing and Speaking Modules.

**Table 3 Mean, standard deviation and standard error of measurement for Writing and Speaking Modules**

| | Mean | SD | SEM |
|---|---|---|---|
| Academic Writing | 5.70 | 1.03 | 0.46 |
| General Training Writing | 5.73 | 1.12 | 0.50 |
| Speaking | 6.05 | 1.10 | 0.47 |

The SEM derived from the reliability figures should be interpreted in terms of the final band scores reported for Writing and Speaking (which are only reported as whole bands).

Experimental generalisability studies were also carried out as part of the IELTS Speaking Revision Project (1998–2001) and the IELTS Writing Revision Project (2001–2004). The study conducted for the Speaking Revision produced an inter-rater correlation of 0.77, and a g-coefficient of 0.86 for the operational single-rater condition (see Shaw 2004); the Writing Revision study produced an inter-rater correlation of 0.77 and g-coefficients of 0.85–0.93for the operational single-rater condition (see Taylor & Jones 2001).

Performance of test materials in the Writing and Speaking Modules is routinely analysed to check on the comparability of different test versions and to ensure any variation is within the acceptable limit of 0.5 of a band. Mean band scores for the Academic Writing versions released in September 2004, and for which a sufficient sample size has been obtained, ranged from 5.41 to 6.13. Mean band scores for the General Training Writing versions released in September 2004 ranged from 5.59 to 6.00.

**Table 4: Mean band scores for whole population**

| MODULE | Listening | Reading | Writing | Speaking | Overall |
|---|---|---|---|---|---|
| Academic | 6.03 | 5.90 | 5.75 | 6.04 | 6.00 |
| General Training | 5.84 | 5.62 | 5.87 | 6.12 | 5.93 |

**Table 5: Mean band scores by gender**

| MODULE | Listening | Reading | Writing | Speaking | Overall |
|---|---|---|---|---|---|
| **Female** | | | | | |
| Academic | 6.10 | 5.97 | 5.83 | 6.13 | 6.08 |
| General Training | 5.84 | 5.58 | 5.92 | 6.11 | 5.93 |
| **Male** | | | | | |
| Academic | 5.96 | 5.84 | 5.67 | 5.99 | 5.93 |
| General Training | 5.84 | 5.64 | 5.85 | 6.13 | 5.93 |

Mean band scores for Speaking versions released in September 2004 ranged from 5.87 to 6.35.

## Test-taker performance

IELTS is assessed on a nine-band scale and reports scores both overall and by skill Module. Overall Band Scores for Academic and General Training candidates in 2004 are reported here for all candidates and by gender (Tables 4 and 5) together with mean band scores for the most frequent first languages for the individual skill Modules (Tables 6, 7,8).

General Training candidates showed greater competence in their Writing and Speaking skills relative to their skills in Reading and Listening. On average, Academic candidates showed less variation across the skills, but find the Writing Module most challenging.

Just over three-quarters of candidates (77.3%) took the Academic Reading and Writing Modules of IELTS and just under a quarter (22.7%) took the General Training Reading and Writing Modules. Overall, the IELTS candidature during the year was 47.1% female and 52.9% male.

Of candidates taking the Academic Modules 49.5% were Female and 50.5% male; 38.9% of candidates taking the General Training Modules were female and 61.1% were male.

The data presented here are also available on the IELTS website (www.IELTS.org) along with previous performance data for IELTS which can be found in the relevant annual reviews and in *Research Notes* 18.

**References**

Shaw, S (2004) IELTS Writing: revising assessment criteria and scales (Phase 3), *Research Notes* 16, 3–7.

Taylor L & Jones, N (2001) Revising the IELTS Speaking Test, *Research Notes* 4, 9–12.

**Table 6: Mean band scores for most frequent first languages (Academic)**

| | Listening | Reading | Writing | Speaking | Overall |
|---|---|---|---|---|---|
| Chinese | 5.65 | 5.75 | 5.34 | 5.49 | 5.64 |
| Tagalog | 6.35 | 6.07 | 6.30 | 6.63 | 6.40 |
| Urdu | 6.15 | 5.76 | 5.89 | 6.08 | 6.04 |
| Bengali | 5.41 | 5.32 | 5.37 | 5.72 | 5.52 |
| Malayalam | 6.21 | 5.92 | 6.16 | 6.29 | 6.21 |
| Arabic | 5.92 | 5.65 | 5.62 | 6.27 | 5.93 |
| Hindi | 6.85 | 6.34 | 6.47 | 6.80 | 6.68 |
| Korean | 5.87 | 5.73 | 5.37 | 5.67 | 5.73 |
| Telegu | 6.44 | 5.87 | 6.15 | 6.27 | 6.25 |
| Punjabi | 6.22 | 5.72 | 5.94 | 6.12 | 6.07 |
| Thai | 5.72 | 5.66 | 5.18 | 5.56 | 5.60 |
| Japanese | 5.84 | 5.71 | 5.40 | 5.75 | 5.74 |
| Gujurati | 6.40 | 5.80 | 5.89 | 6.09 | 6.10 |
| Tamil | 6.57 | 6.14 | 6.19 | 6.51 | 6.41 |
| Farsi | 5.73 | 5.74 | 5.81 | 6.28 | 5.97 |
| Spanish | 6.29 | 6.52 | 6.08 | 6.63 | 6.44 |
| Indonesian | 6.11 | 6.05 | 5.38 | 5.78 | 5.89 |
| Vietnamese | 5.65 | 5.82 | 5.63 | 5.68 | 5.76 |
| Malay | 6.78 | 6.49 | 6.01 | 6.34 | 6.46 |
| German | 7.31 | 7.11 | 6.84 | 7.30 | 7.21 |

**Table 7: Mean band scores for most frequent first languages (General Training)**

| | Listening | Reading | Writing | Speaking | Overall |
|---|---|---|---|---|---|
| Chinese | 5.67 | 5.78 | 5.62 | 5.68 | 5.76 |
| Punjabi | 5.95 | 5.45 | 6.02 | 6.14 | 5.95 |
| Urdu | 5.74 | 5.36 | 5.87 | 6.11 | 5.84 |
| Hindi | 6.44 | 5.94 | 6.41 | 6.71 | 6.44 |
| Arabic | 5.42 | 5.18 | 5.40 | 5.99 | 5.56 |
| Malayalam | 5.62 | 5.22 | 5.78 | 5.80 | 5.47 |
| Gujurati | 5.87 | 5.34 | 5.67 | 5.85 | 5.74 |
| Tagalog | 6.11 | 5.77 | 6.25 | 6.41 | 6.20 |
| Korean | 5.42 | 5.42 | 5.23 | 5.38 | 5.43 |
| Spanish | 5.62 | 5.97 | 5.90 | 6.24 | 6.00 |
| Russian | 5.53 | 5.70 | 5.68 | 5.89 | 5.77 |
| Bengali | 5.59 | 5.38 | 5.76 | 6.16 | 5.79 |
| Farsi | 5.27 | 5.20 | 5.64 | 6.01 | 5.61 |
| Tamil | 6.04 | 5.69 | 5.97 | 6.24 | 6.05 |
| Japanese | 5.47 | 5.40 | 5.26 | 5.55 | 5.49 |
| Singhalese | 5.80 | 5.45 | 5.81 | 6.11 | 5.86 |
| Telegu | 6.19 | 5.65 | 6.09 | 6.38 | 6.15 |
| Marathi | 6.38 | 5.89 | 6.44 | 6.69 | 6.41 |
| Indonesian | 6.11 | 5.94 | 5.61 | 5.90 | 5.97 |
| German | 6.70 | 6.48 | 6.56 | 6.96 | 6.74 |

**Table 8: Mean band scores for most frequent countries of origin (Academic)**

| | Listening | Reading | Writing | Speaking | Overall |
|---|---|---|---|---|---|
| China | 5.50 | 5.61 | 5.23 | 5.40 | 5.51 |
| India | 6.51 | 6.05 | 6.24 | 6.45 | 6.38 |
| Philippines | 6.35 | 6.08 | 6.31 | 6.64 | 6.41 |
| Pakistan | 6.05 | 5.71 | 5.85 | 6.02 | 5.98 |
| China (Hong Kong SAR) | 6.42 | 6.56 | 5.92 | 6.04 | 6.30 |
| Bangladesh | 5.30 | 5.22 | 5.28 | 5.63 | 5.42 |
| Korea, South | 5.87 | 5.73 | 5.38 | 5.68 | 5.73 |
| Malaysia | 6.82 | 6.60 | 6.11 | 6.39 | 6.54 |
| Thailand | 5.72 | 5.66 | 5.18 | 5.57 | 5.60 |
| Japan | 5.84 | 5.71 | 5.40 | 5.75 | 5.74 |
| Taiwan | 5.60 | 5.58 | 5.24 | 5.64 | 5.59 |
| Iran | 5.74 | 5.74 | 5.81 | 6.28 | 5.97 |
| Indonesia | 6.11 | 6.05 | 5.39 | 5.79 | 5.90 |
| Vietnam | 5.65 | 5.82 | 5.63 | 5.67 | 5.76 |
| Sri Lanka | 6.36 | 5.93 | 6.01 | 6.46 | 6.26 |
| Germany | 7.30 | 7.12 | 6.81 | 7.29 | 7.20 |
| Nigeria | 6.48 | 6.43 | 7.21 | 7.65 | 7.00 |
| United Arab Emirates | 5.75 | 5.10 | 5.49 | 5.98 | 5.65 |
| Russia | 6.44 | 6.34 | 6.10 | 6.74 | 6.47 |
| Nepal | 6.12 | 5.74 | 5.60 | 5.76 | 5.86 |

**Table 9: Mean band scores for most frequent countries of origin (General Training)**

| | Listening | Reading | Writing | Speaking | Overall |
|---|---|---|---|---|---|
| India | 6.09 | 5.61 | 6.09 | 6.28 | 6.08 |
| China | 5.63 | 5.74 | 5.61 | 5.62 | 5.72 |
| Pakistan | 5.67 | 5.31 | 5.84 | 6.04 | 5.78 |
| Philippines | 6.12 | 5.77 | 6.27 | 6.43 | 6.21 |
| Korea, South | 5.42 | 5.41 | 5.23 | 5.38 | 5.43 |
| Iran | 5.27 | 5.19 | 5.63 | 6.01 | 5.61 |
| Bangladesh | 5.43 | 5.24 | 5.63 | 6.05 | 5.65 |
| Sri Lanka | 5.82 | 5.45 | 5.82 | 6.13 | 5.87 |
| Japan | 5.48 | 5.40 | 5.27 | 5.55 | 5.49 |
| Russia | 5.56 | 5.79 | 5.72 | 5.94 | 5.83 |
| China (Hong Kong SAR) | 5.93 | 6.06 | 5.77 | 5.86 | 5.99 |
| Indonesia | 6.11 | 5.93 | 5.61 | 5.90 | 5.97 |
| Taiwan | 5.44 | 5.63 | 5.46 | 5.73 | 5.63 |
| Jordan | 5.64 | 5.39 | 5.57 | 6.11 | 5.75 |
| Colombia | 5.08 | 5.52 | 5.50 | 5.88 | 5.56 |
| United Arab Emirates | 4.61 | 3.99 | 4.41 | 5.27 | 4.63 |
| Ukraine | 5.50 | 5.65 | 5.73 | 5.96 | 5.77 |
| Romania | 5.59 | 5.76 | 5.86 | 6.17 | 5.93 |
| Mexico | 6.13 | 6.37 | 6.23 | 6.70 | 6.43 |
| Malaysia | 6.64 | 6.41 | 6.37 | 6.75 | 6.60 |

# IELTS award news

We announce below the winner of the IELTS Masters Award 2005, the recipients of the 11th round of the Joint-funded Research Program, followed by the call for entries to the 2006 Masters Award.

### Winner of the IELTS Masters Award 2005

In 1999, the three IELTS partners – the University of Cambridge ESOL Examinations, The British Council, and IDP: IELTS Australia – inaugurated the IELTS MA Thesis Award, an annual award of £1000 for the masters level thesis or dissertation in English which makes the most significant contribution to the field of language testing. Since 1999, there have been 4 winners of the award – from Canada, Australia and USA.

For the 2005 Award, now called the IELTS Masters Award, submissions were accepted for masters theses completed and

approved in 2004. The IELTS Research Committee, which comprises members of the three partner organisations, met in November 2005 to review the shortlisted submissions and the Committee was once again impressed with the quality of work received. After careful consideration, the Committee decided to announce one winner: Fumiyo Nakatsuhara – for her thesis entitled 'An Investigation into Conversational Styles in Paired Speaking Tests'. Fumiyo completed her thesis at the Department of Language and Linguistics, University of Essex (UK) and her supervisor was Anthony D Lilley.

Fumiyo's full abstract appears below:

> As studies of oral testing have suggested a great desirability for the introduction of paired testing, a paired format, where non-native test-takers are paired and examined together, has recently become a popular tool to access oral interactional communication ability. This introduction, however, has also caused concern with regard to how the pairing of test-takers should appropriately be conducted, since paired interlocutor characteristics could be significant variables influencing one's performance. Particularly, examination boards and some teachers seem to consider the proficiency of interlocutors as the most important factor in determining performance on the paired tests, despite the fact that little is actually understood about the effect. Therefore, the present study addresses the following research questions:
>
> 1) *Are conversational styles of dyads different between same-proficiency level pairs (SPL pairs) and different-proficiency level pairs (DPL pairs)?*
> 2) *Are dyadic interactions with different-ability speakers asymmetrical? If so, to what extent are they asymmetrical? How are they asymmetrical?*
>
> Data were collected from 12 sessions of SPL pairs and 12 sessions of DPL pairs in which advanced and intermediate learners of English simulated the collaborative part of the Cambridge Certificate in Advanced English (CAE) examination. All the sessions were video-taped and transcribed following Conversation Analysis (CA) conventions, and then both quantitatively and qualitatively analysed about three components of conversational styles: *interactional contingency, goal-orientation and quantitative dominance.* The result revealed that while the proficiency of the paired interlocutor may have a small impact on the element of interactional contingency, there are many more similarities than differences between SPL and DPL pairs. Moreover, although the examination into DPL pairs suggests that advanced candidates may be slightly more goal-orientated and speak more than paired intermediate partners, the level of variability was not big enough to differentiate interactional styles of SPL and DPL pairs. Additionally, this research discovered that some candidates help their partners in their dyadic interactions, and the accommodative behaviours could contribute to the balanced conversational styles produced. All in all, the presence of different proficiency levels in paired tests may not be as serious a concern as anticipated, and the findings encourage us to continue to introduce the paired format as a useful mode of oral assessment.

The Research Committee noted that Fumiyo's dissertation was an excellent piece of research. Of interest to test designers and researchers in the field, her work should constitute the basis of further research in this area. In terms of the importance of the topic, sound rigorous scholarship, the appropriateness and logic of the conclusions drawn or recommendations made, contribution to knowledge, clearly stated research questions and an appropriate research design, comprehensive review of relevant literature, with correctly interpreted references to other authors and works, this dissertation was judged to be a worthy winner of the 2005 award.

Fumiyo will be presented with her award (a cheque and certificate) at the 28th Annual Language Testing Research Colloquium (LTRC), June 29–July 1 2006 at the University of Melbourne, Australia. For further information about LTRC see http://www.languages.unimelb.edu.au/ltrc2006/

### Recipients of Joint Research Program funding (Round 11)

Once again competition was intense for the funding available under the 11th round of the Joint-funded Research Program. The IELTS Research Committee selected the following projects for research funding in 2005–6:

**Studies funded under round 11 of the IELTS Joint-funded Research Program**

**Round 11 / 2005**

| Topic | Researchers |
|---|---|
| A cognitive validation of the lecture-listening component of the IELTS listening paper | Dr John Field, UK |
| Investigating IELTS exit score gains in higher education. | Dr Kieran O'Loughlin & Dr Sophie Arkoudis, The University of Melbourne, Australia |
| An impact study into the use of IELTS as an entry criterion for professional associations – USA, New Zealand and Australia. | Ms Glenys Merrifield, GBM & Associates, Australia |
| The effect of memorized learning on the writing scores of Chinese IELTS test takers. | Dr Christine Pegg & Prof Alison Wray, University of Cardiff, UK |
| The IELTS General Training Module as a predictor of performance in practical tertiary programmes. | Dr Hilary Smith & Prof Stephen Haslett, Systemetrics Research Limited, New Zealand |
| Accessibility of IELTS GT Modules to 16/17 year old students studying English in selected Asian countries. | Ms Jan Smith, Australia |
| The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in the first year of their courses at a British university. | Prof Cyril Weir, Luton University/Luton Business School, UK |

See *Research Notes* issues 8 and 20 for a list of previous recipients of funding. See the Research section of the IELTS website for a list of research reports published by IELTS Australia.

## Call for Entries for IELTS Masters Award 2006

Each year the IELTS partners sponsor an award of £1000 for the Masters level dissertation or thesis which makes the most significant contribution to the field of language testing. The entry procedures and timetable for the 2006 award are given on the following page.

**Submission and evaluation procedures**

Dissertations will only be considered eligible if they were submitted and approved by your university in 2005. Dissertations completed in 2006 will not be considered eligible for the 2006 award but may be submitted the following year. Submissions should be for dissertations written in partial or total fulfilment of the requirements for a Masters degree or its equivalent.

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

Dr Lynda Taylor/Stuart Shaw
Research and Validation Group
University of Cambridge ESOL Examinations
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

The IELTS Research Committee will review the submissions and shortlist potential award winners. For all shortlisted dissertations a full copy of the dissertation will be requested and a further reference may be sought. Shortlisted dissertations will be reviewed and evaluated by the IELTS Research Committee according to the following criteria:

- Rationale for the research
- Contextualisation within the literature
- Feasibility of outcomes
- Design of research question(s)
- Choice and use of methodology
- Interpretation and conclusions
- Quality of presentation
- Use of references
- Contribution to the field
- Potential for future publication.

The Committee's decision is final.

**Timetable**

The following timetable will apply in 2006:

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

Please note that submission details may change from year to year and it is therefore important that the most current procedures are consulted. Details of the application process for the IELTS Masters Award 2006 can also be found on the IELTS website – www.ielts.org

# ESOL Special Circumstances 2004: a review of Upper Main Suite provision

**MIKE GUTTERIDGE**, CAMBRIDGE ESOL CONSULTANT

## Introduction

This review presents a selection of elements from the Cambridge ESOL Special Circumstances Report for 2004. The Special Circumstances Report is prepared on an annual basis and provides a general survey of work carried out, together with a detailed statistical analysis of cases dealt with. The focus of this paper is mainly on FCE, CAE and CPE (the Upper Main Suite of Cambridge ESOL examinations) as the majority of cases dealt with by Cambridge ESOL involve these examinations.

Special Circumstances covers three main areas: special arrangements, special consideration, and cases of malpractice in respect of Cambridge ESOL products. These three areas may be defined more precisely as follows:

**Special Arrangements:** are made for candidates with special requirements before an examination is taken so that, as far as possible, they are then able to take the examination on an equal footing with other candidates. For example, candidates with a permanent disability, such as hearing/sight impairment, dyslexia or a speech impediment; or short-term difficulties (for example, a broken arm) may need arrangements such as modified papers, readers or amanuenses, or extra time.

**Special Consideration:** is given to candidates who are affected immediately before or during an examination by adverse circumstances. Examples include illness, bereavement or circumstances affecting the conditions under which an exam is taken. Special Consideration is applied for after the candidate sits an examination.

**Malpractice:** (defined as any conduct which has the intention, or effect, of giving an unfair advantage to one or more candidates) is brought to the attention of Cambridge ESOL via reports from Centres, reports from examiners or inspectors, and through routine statistical checks applied to candidates' answers.

**Table 1: Special circumstances candidates for all Cambridge ESOL products 2001–4**

| Year | Special Arrangements (candidates) | Special Consideration (candidates) | Malpractice Cases (each case may involve one or more candidates) |
|---|---|---|---|
| 2000 | 948 | 6441 | 120 |
| 2001 | 1135 | 11646 | 122 |
| 2002 | 1355 | 10655 | 169 |
| 2003 | 2209 | 12950 | 245 |
| 2004 | 1443 | 10270 | 261 |

Table 1 provides a snapshot of number of cases over the past five years and lists total numbers of candidates applying for special arrangements, special consideration and those involved in cases of malpractice for all Cambridge ESOL products.

Although these numbers may appear small in comparison to the total test-taking population, it should be remembered that all the above cases are dealt with on an individual basis. For example, a blind candidate may require separate facilities for taking the examination, a specially modified question paper, an individual invigilator and/or reader or amanuensis on the day, and extra time to complete their papers. The same need for individual (and often lengthy) attention also applies to applications for special consideration and, of course, in all cases where candidates have been reported for malpractice.

## Special Arrangements

Numbers of applications for special arrangements for all candidates taking FCE, CAE and CPE in all categories (including extra time, etc.) in 2000–2004 are shown in Table 2.

**Table 2: Applications for special arrangements for UMS candidates 2000–4**

| Year | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|
| Candidates | 670 | 966 | 1355 | 1948 | 1149 |

Table 3 gives a brief overview of important categories of special arrangements for FCE, CAE and CPE in 2003 and 2004.

**Table 3: Key categories of special arrangements for UMS candidates in 2003–4**

| Main Categories | 2003 | 2004 |
|---|---|---|
| Braille versions of papers | 21 | 20 |
| Enlarged print papers | 50 | 41 |
| Lip-reading versions of listening papers | 31 | 38 |
| Special versions of listening papers | 85 | 90 |
| Separate marking of writing papers | 950 | 254 |
| Exemption for listening or speaking components | 22 | 24 |

We can note the following about all categories in Table 3.

*Braille versions*

Approximately the same number of blind candidates applied for Braille versions of FCE, CAE and CPE in 2004. The majority of applications were for uncontracted Braille.

*Enlarged print versions*

For some exams, these are supplied as either A3 or A4 versions. In the case of the former, the standard paper is enlarged to A3, giving a typical font size of 15.5; in the latter, font size, layout, etc, are standardised, with the font typically being 18 point bold. For most types of visual impairment, therefore, the A4 version, where available, is more appropriate than the A3.

Numbers of partially-sighted candidates applying for these versions of FCE, CAE and CPE fell in 2004. In 2004, the majority of applications were for A3 format question papers, possibly as the result of a misunderstanding as to which version results in larger print.

*Hearing-impaired (lip-reading) versions*

An increase in numbers applying for special lip-reading versions of FCE, CAE and CPE Listening Tests was noted in 2004.

*Special versions of listening papers*

Approximately the same number of candidates applied for these specially recorded versions as in 2003. They are available to blind, partially-sighted, physically disabled, and temporarily disabled candidates (ie: in any circumstance where a candidate is unable to write notes/answers while they are listening).

*Separate marking of writing papers*

As in previous years, candidates with Specific Learning Difficulties (eg: dyslexia) were able to apply to have their written work marked 'separately' (ie: with spelling errors being disregarded) in 2004. Separate marking of writing papers was discontinued from December 2004, following research into the most appropriate arrangements for candidates with Specific Learning Difficulties taking Cambridge ESOL examinations.

*Exemption for listening or speaking components*

Relatively small numbers of candidates applied for exemption from FCE, CAE or CPE Listening or Speaking components in 2004, but there were slightly more candidates than in 2003. This arrangement can prove particularly useful in cases where candidates have severe speaking or listening difficulties. Candidates applying for this arrangement receive a certificate endorsement or 'indication', if an overall passing grade is achieved.

## Special Consideration

Overall, it is interesting to note that the total number of applications for Special Consideration for all ESOL examinations processed by Cambridge ESOL decreased by 19% between 2003

and 2004. Applications for special consideration are received for a wide variety of reasons. Candidates may have been affected before an examination by personal illness, accident or bereavement; alternatively, they may have been affected by adverse circumstances during the actual taking of the examination, e.g. unexpected noise, equipment failure, or some other disruption.

Action taken depends on the type and severity of the circumstances reported, although appropriate action to compensate candidates affected is only possible if the nature of the problem is accurately and comprehensively described by examination supervisors and staff at the test centre.

Table 4 gives a comparison of total numbers of applications (numbers of candidates) in 2000–2004 for FCE, CAE and CPE.

**Table 4: Applications for special consideration for UMS candidates 2000–4**

| Year | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|
| Candidates | 5619 | 11046 | 8438 | 10265 | 9278 |

There was a 10% decline in applications for special consideration for FCE, CAE and CPE candidates between 2003 and 2004.

## Malpractice

Malpractice may be reported by centres, examiners or detected by routine statistical checks applied to candidates' answer sheets. The Cambridge ESOL Malpractice Committee scrutinises all reports and investigates where appropriate. A fine balance is required in dealing with cases of malpractice, where the circumstances of each report of irregular conduct must be investigated fairly and carefully whilst ensuring that the integrity of a particular examination is preserved.

In 2004, 261 malpractice cases were dealt with for all Cambridge ESOL products. Each 'case' may involve one or more candidates. The total number of cases for 2003 was 245.

Malpractice procedures apply to all Cambridge ESOL products, including, for example, Young Learners English Tests. Teachers are asked, via Examination Reports and other documents, to remind Young Learners candidates that the same rules apply to young children taking Young Learners tests as to any other candidates.

## Modified paper production

In 2004, as in previous years, modified versions of question papers for candidates with sight or hearing difficulties were prepared. At each session, question papers and other material are initially reviewed for content suitability, before any modifications or adaptations are attempted. Following this content review, material is prepared for brailling.

Table 5 shows the total number of modified materials for FCE, CAE and CPE examinations (including question papers, recordings and modified speaking tests) produced for use in 2004. Supervisor's Booklets (instructions and tapescript for administering special needs versions of listening tests, including hearing-impaired versions) are produced routinely for each listening test modified.

**Table 5: Number of modified materials for UMS examinations 2004**

| Modification | 2004 |
|---|---|
| Braille versions | 54 |
| Enlarged print versions (A4) | 54 |
| Special recordings | 7 |
| Supervisor's Booklets | 7 |
| Hearing-impaired versions | 7 |
| Modified Speaking Tests | 9 |

Modified Braille and A4 Enlarged Print versions of FCE, CAE and CPE are carried out routinely for each session with the expectation that there will be at least one application for each version per syllabus. Following modifications, print versions of Brailled material are produced, and brailling is requested when an application is received from a Centre.

## Conclusion

This review has highlighted the scope of work carried out by Cambridge ESOL and the volume of applications for special arrangements and special consideration relating to Upper Main Suite examinations in 2004. The range of modifications carried out is subject to continuous review to ensure that examination material is made as accessible as possible to all candidates, without compromising assessment objectives.

The area of Special Circumstances is complex precisely because such a fine balance is required between allowing candidates with special requirements arrangements which enable them to be placed on an equal footing with other candidates but not advantaging them to the extent that the assessment objectives of a particular examination are compromised.

# Current research and development activities

## Computer-based testing

The first administration of the new computer-based version of PET was held in November 2005 in a number of centres in Europe. Feedback from the centres involved was very positive. Staff at these centres were happy with the ease of administration of the test and many said that they would recommend CB PET to other centres. They also said that candidates were very happy with the test experience, but expressed a desire for more practice materials. A sample test is now available from Cambridge ESOL on CD-ROM and both Longman and Cambridge University Press are developing computer-based practice test material for publication in 2006.

Recent research included a plurilingual study to investigate the comparability of CB BULATS English, French, German and Spanish versions. Each participant took at least two different language versions of CB BULATS, as well as completing can-do self assessments for each language. This design allows the variability that may result from candidates under- or over-estimating their ability compared with other candidates to be accounted for, assuming that a candidate would be equally lenient or harsh on their self-assessment in each language. The findings indicated that for the English version the candidates' self-assessments were higher compared to their score in the test than for the French, German and Spanish versions (in that order). This might be taken as an indication that candidates tend to find the English version the most difficult of the CB BULATS languages, or could equally be because participants were more likely to overestimate their ability in English. Due to the difficulties in finding suitable participants for this kind of study, the sample was rather limited. Plurilingual data will therefore continue to be collected on an ongoing basis, so that a larger scale investigation can be conducted in future.

Cambridge ESOL held a Development Projects EXPO on the 30th November 2005 to enable some of the technological developments currently being worked on to be showcased internally. The day was well attended by staff throughout the Cambridge Assessment Group, including chief executive Simon Lebus, and consisted of a series of demonstrations and presentations, each focusing on different websites, products and services. Paul Seddon began the day with a presentation of the Online Test Delivery System, which will be used to deliver a range of e-assessments directly to centres online, enabling shorter lead-in times and turn-around times and greater frequency of sessions. This presentation was particularly timely, since the system was used to deliver live assessments for the first time during November, with the launch of CB PET. Mitra Assadi then gave an overview of how User Acceptance Testing is carried out within Cambridge ESOL to ensure our computer-based products and systems are of the highest quality before they are released into live use. Peter Simmons followed with a talk on Formative Assessment, including a discussion of how technology might be used to integrate assessment into learning and help learners and teachers to set and monitor their learning aims. Andrew Milbourn ended the day of presentations with a description of the developments which have taken place in the last year to improve Cambridge ESOL's internal information systems. A wide range of demonstrations also took place throughout the day, which allowed visitors the opportunity to experience hands-on interaction with the systems. These included developments such as the ESOL online entries system, the new candidate results website, digital object marking software (which will be used to enable candidate responses to be marked on-screen as part of the Electronic Script Management project) and ePortfolios, as well as products such as CB PET, CB BULATS and CBIELTS. The event was considered very useful and staff enjoyed the opportunity to see some of our new developments in action.

## Asset Languages

The last update on Asset Languages commented on the contribution that pretesting has made to establishing the vertical link between Asset Languages stages. Recently pretests that are made up of two adjacent stages have also been used in order to improve the quality of the vertical linking. Cross-stage pretests have thus far been taken by candidates in Chinese and Japanese. Further cross-stage pretests are being developed for Panjabi, Urdu, Italian, French, German and Spanish and will become part of the routine pretesting operation.

The Asset Languages Research Agenda has been established and a set of research projects have been initiated. Three projects focus on cross-language issues and will use learner-centred can-do methods. One project will look at learners of both Chinese and French and ask learners to rate their relative ability in these two languages across the four skills of listening, reading, writing and speaking. As Asset Languages is designed for both community and modern foreign language learners, pretest data, first language information and estimates of National Curriculum levels will also be used in the analysis to learn more about candidates and their ability. Further development of methodologies for cross-language standardisation and equating will feed into work relating Asset Languages assessments to the Common European Framework of Reference.

An analysis of the performance of task types has been undertaken using data from the May 2005 pilot administrations. The analysis highlights similarities in task type performance across languages, linking this to the task construct, and identifies whether particular task types perform well or badly with regard to their difficulty and fit (similar to discrimination). A report focusing on Breakthrough stage tasks has been produced to see how well the Asset Languages aims for this level are being met. Issues such as syllabus dependency, intended and actual strategies for completing tasks are covered.

An evaluation, reviewing the success of the scheme to date in terms of its rationales of motivating learners and accrediting language proficiency, has been completed for the Qualifications and Curriculum Authority (QCA) in the UK.

For more information on Asset Languages visit www.assetlanguages.org.uk

## International Legal English Certificate 2005 trial

The International Legal English Certificate (ILEC) is a certificated examination that is being developed by Cambridge ESOL in collaboration with Translegal, a company of lawyer-linguists. ILEC is aimed at law students and practising lawyers who are seeking employment in an international legal setting and wish to obtain a law-related English language qualification. ILEC will have four papers, Reading, Writing, Listening and Speaking, and is aimed at candidates at B2 and C1 on the Common European Framework of Reference. It is intended that the first administration of ILEC will take place in selected countries in May 2006.

A trial of the examination was carried out in May 2005. Its purpose was to provide evidence of the reliability and validity of ILEC and to provide test designers with key information on tasks and the format of the test. Three hundred and twenty candidates from 12 centres in Europe, covering different jurisdictions, completed Reading, Listening and Writing trial papers. They also completed a calibrated general English test and questionnaires on their perceptions of ILEC. Information on age, qualifications, experience in working in a legal environment and so on was also collected.

From the candidate questionnaires and shadowing of trial candidates taking Reading and Writing papers it was found that a reasonable amount of time has been provided for the majority of the test-takers to be able to complete the components. Initial investigations into the reliability of the test papers indicated that the examination would accurately assess candidates' level of proficiency at B2 and C1.

The trial revealed evidence that ILEC has a substantial degree of face validity in that the majority of candidates and instructors believe the topics, texts and language are similar to what they may be expected to meet in their working legal environment. For example, in Reading and Writing around 65% of candidates agreed or agreed strongly that the topics, texts and language are authentic. In Listening, around 55% of candidates agreed or agreed strongly that the tasks are authentic and 27% of candidates showed no preference. While response bias cannot be discounted, these figures provide encouraging evidence of the face validity of ILEC. They also provide evidence of situational authenticity in that a majority of candidates and instructors believe that ILEC topics, texts and language mirror those in the target language use situation.

Further research into ILEC will continue to focus on the application of Weir's (2005) Socio-Cognitive Framework for Developing and Validating Tests to ILEC (see also Weir and Shaw 2005). This framework, through a systematic description of attributes of the test-taker, the test format and scoring methods, allows us to investigate and provide evidence of the interactional authenticity of the test from a number of perspectives. For example, the research on speededness, noted above, suggests that, as in the target language use situation, candidates have been provided with a reasonable amount of time to complete tasks to their best ability.

For more information on ILEC visit www.legalenglishtest.org

**References**

Weir, C (2005) *Language Testing and Validation: an evidence-based approach*, Oxford: Palgrave.

Weir, C and Shaw, S (2005) Establishing the Validity of Cambridge ESOL Writing Tests: towards the implementation of a socio-cognitive model for test validation, *Research Notes* 21, 10–14.

# Conference reports

2005 proved to be a busy year for Cambridge ESOL staff colleagues in terms of presenting at national and international conferences. In the previous two issues of *Research Notes* we reported on symposia and papers given at the ALTE (Berlin) and AILA (USA) conferences in May and July 2005. Another key conference for us was the 27th Annual Language Testing Research Colloquium (LTRC) in Ottawa, which is reported on below, along with BAAL (Bristol), ALTE (Cardiff) and LTF (Cambridge, UK).

## LTRC 2005

The theme for this year's LTRC conference was 'Challenges, issues, impacts: The interplay of research and language testing practice'. Cambridge ESOL staff contributed a workshop and two papers to this event which took place from 18–22 July at the University of Ottawa in Ontario, Canada.

One of the two pre-conference workshops at LTRC was led by Prof Anne Lazaraton (University of Minnesota) and Dr Lynda Taylor (Cambridge ESOL); their workshop was entitled 'Qualitative research methods in language test development and validation'. This one-day event aimed to introduce participants to the underlying premises of qualitative research, particularly its relevance in the context of language testing and assessment. After a brief overview of some noteworthy research in this field using qualitative research methods, the presenters examined the application of qualitative methodologies to speaking and writing tests where such methods have proved particularly fruitful over the past 10–15 years. Discussion and worked practice activities were used to explore how qualitative analysis can provide us with rich insights into the behaviours (in terms of both *process* and *product*) of test-takers, interlocutors and raters in speaking and writing assessment. The workshop made extensive reference to qualitative studies conducted on the direct speaking/writing tests in our Cambridge ESOL examinations; much of this work has been published in volumes in the *Studies in Language Testing* series or reported in *Research Notes*.

Around 60 participants from all over the world attended the

workshop. Many were graduate students or 'researchers-in-training', though more experienced researchers welcomed the chance for a 'refresher course'. Hopefully, the sessions helped develop the confidence and skills of all participants so that they feel better equipped to conduct their own qualitative research studies in the field of language testing and assessment. Articles based on the workshop are due to appear in the LTRC 2005 Proceedings and in the journal *Language Assessment Quarterly.*

### Reporting on standard-setting studies for IELTS

The opening up of international borders and the growth in global employment opportunities has led to considerable expansion in the number of overseas health professionals (e.g. doctors, nurses) entering the UK, USA and other English-speaking countries to work in these nations' health services. Not surprisingly, internationally recognised English proficiency tests (such as IELTS and TOEFL) are increasingly being used as part of the recruitment process to provide information about the language proficiency of health professionals for whom English is a second language – especially those who are seeking official registration or license to practice. Two papers were presented at LTRC 2005 reporting on recent standard-setting studies conducted with IELTS; both studies sought to identify the minimum level of English language proficiency needed for specific health professional groups to work safely and effectively, and to determine the associated cut scores on the test.

The first paper was presented by Jay Banerjee (Lancaster University) and Lynda Taylor (Cambridge ESOL), entitled 'Setting the standard: what English language abilities do overseas trained doctors need?' It described a review of the UK General Medical Council's (GMC) minimum language proficiency criteria for International Medical Graduates (IMGs), taking into account the judgements of three key stakeholder groups: patients, doctors, and other health workers (such as nurses or physiotherapists). These represent the main groups that come into contact with the IMGs. The paper described the study design and presented the findings, including recommended minimum IELTS writing and speaking scores for consideration by the GMC. It discussed the differences between the stakeholder groups in their perceptions of what constitutes adequate language proficiency. This study demonstrated the value of conducting standard-setting studies with all stakeholder groups and raised important questions about how language testers should reconcile differences in stakeholder opinion.

A second paper on a similar theme was presented by Lynda Taylor (Cambridge ESOL) and Thomas O'Neill (National Council of State Boards of Nursing – NCSBN, USA) called 'Safe to practise? Setting minimum language proficiency standards among nursing professionals'. This paper reported on a US-initiated study to identify the minimum level of English language proficiency needed for entry-level nursing professionals to work safely and effectively, and to determine the appropriate band scores on IELTS. The paper described the different methodologies adopted for cut score estimation consistent with the characteristics of each language subtest: a) a modified Angoff (1971) method for the Reading and Listening subtests, and b) a modified Analytical Judgement Method (Plake and Hambleton, 2000) for the Writing and Speaking subtests. The selection and qualifications of the panellists were outlined, along with the training given to them. Findings were reported together with recommendations for policy decisions.

These two papers illustrated in practice how standard-setting studies can be designed and conducted and how the outcomes can be used to assist policy makers in the setting of reasonable and defensible cut scores. In addition, the two studies demonstrate how positive collaboration can be achieved between key test stakeholders, i.e. professional bodies who elect to use the tests (e.g. GMC, NCSBN), independent assessment experts commissioned by the test user to research the tests (e.g. Lancaster University, UK, BUROS Institute of Mental Measurements, USA), and the actual producers of the test (Cambridge ESOL and the other IELTS partners). A paper based on the second study reported above has recently been submitted to the journal *Language Assessment Quarterly* and is currently undergoing review.

## BAAL 2005

In September 2005, Cambridge ESOL staff were once again involved in delivering papers at the annual conference of the British Association of Applied Linguistics (BAAL) in Bristol, UK. The theme of this year's conference was 'Language, Culture and Identity in Applied Linguistics'. Cambridge ESOL was also the major sponsor of this event.

### 'Native' or 'non-native' speaker? A question of language, culture or identity?

This colloquium was convened by Lynda Taylor and included contributions from acknowledged experts in the fields of applied linguistics, language pedagogy and language assessment with a shared interest in the linguistic variety and issues of NS/NNS-hood and use: Prof Alan Davies (University of Edinburgh); Prof Anne Lazaraton (University of Minnesota); Prof Barbara Seidlhofer (University of Vienna); and Prof Janina Brut-Griffler (currently at the University of Buffalo, New York).

For many years the concept of the 'native speaker' (NS) dominated in applied linguistics, especially in language teaching, learning and assessment; the assumption was that the NS is the best teacher, the ideal model for language use, and the natural judge of language ability. However, the early 1990s gave rise to a debate focusing on the validity and usefulness of the traditional NS construct (Davies, 1991; Medgyes, 1994); this debate has intensified over the past 10–15 years due partly to the process of globalisation and the increasing recognition and currency of linguistic varieties.

The colloquium had a three-fold aim: to review current thinking on the NS/NNS distinction as it relates to issues of language, culture and identity; to consider current policy and practice relating to the use of NS/NNS language norms in teaching, learning and assessment, including the status and role of NS/NNSs as teachers and assessors; and to explore how the study of NS/NNS language varieties might offer alternative language models that are both valid and useful, including those which can be informed by recent developments in corpus linguistics. The colloquium was attended by over 70 participants and included opportunities for questions and discussion.

Tony Green also presented on 'Levels of spoken language ability: developing a descriptive common scale' in which he reported ongoing work on the Cambridge Common Scale for Speaking: an evolving frame of reference for the description of spoken language performance designed to provide test users with a clear explanation of the levels addressed by the various Cambridge tests. Tony described how quantitative (test score and corpus) and qualitative (test discourse and participant protocol) analyses of test performance have been combined with insights from stakeholders to inform the ongoing validation and development process. He reported on the operational relevance of the scale for the interpretation of performance at different levels, and the identification of typical performance quality at different levels.

## ALTE 2005

The thirtieth ALTE meeting was held in Cardiff in November 2005. The meeting was hosted by the Welsh Joint Education Committee (WJEC), under the auspices of the UK Presidency of the EU. The conference consisted of an open conference day with the theme of 'Language Assessment for Lifelong Learning' and a series of two day workshops for ALTE Members and Observers held prior to the conference day.

Tony Green conducted a workshop session on issues of quality control in the selection of material for language tests. Drawing an analogy with new car safety tests, he stressed the need to trial or pilot test material before it is used in consequential tests. The session provided an introduction to classical item analysis, demonstrating how statistics can inform decisions to reject, revise or select material for use in a test and offering delegates the chance to review material taken from ALTE tests in the light of statistical information.

Dr Barry O'Sullivan gave a workshop entitled 'Specific purpose testing: Differentiating specific purpose test tasks and general proficiency tests' and offered participants a practical look at generating evidence to show that a test is of a specific purpose, rather than of a more general purpose. This process involved rating relevant specifications on a scale of 1 to 7 (7 = high specificity) and plotting the results for several specifications on the same radar diagram.

Prof Cyril Weir's workshop in Cardiff was based on work by Cambridge ESOL to articulate their approach to assessment in the skill area of Writing. The work builds on Cambridge ESOL's traditional approach to validating tests namely the VRIP approach where the concern was with Validity, Reliability, Impact and Practicality. It explored how the socio-cognitive validity framework described in Weir's *Language Testing and Validation: an evidence-based approach* (2005) might contribute to an enhanced validation framework for use with Cambridge examinations. Weir's approach covers much of the same ground as VRIP but it attempts to reconfigure validity as a unitary concept, and to show how its constituent parts interact with each other. In addition it conceptualises the validation process in a temporal frame thereby identifying the various types of validity evidence that need to be collected at each stage in the test development process. Within each constituent part of the framework criterial individual parameters for distinguishing between adjacent proficiency levels are also identified. Workshop participants were asked to look at examples of writing tasks taken from five different levels of Cambridge ESOL examinations. Through group discussion and interaction, they were asked to indicate where a clear distinction can be made between adjacent levels in terms of a particular contextual parameter.

Neil Jones contributed to a session on the *Council of Europe Pilot Manual for Aligning Examinations to the Common European Framework*, commenting on the benchmarking conferences for speaking organized by the CIEP for French (Sèvres, December 2004) and the Goethe-Institut for German (Munich, October 2005). As a participant in both events and analyst of the Sèvres data, Neil identified some issues for the rating process and illustrated them with findings from the analysis. The assessment criteria used (range, fluency, accuracy, interaction, coherence) were found to be largely redundant, showing that the Common European Framework (CEF) scales did not enable differentiation of subjects beyond the global impression that raters formed of them. This suggested that training effort might be more usefully focused on sharing an understanding of the criterial features of levels.

Three presentations on the conference day related to the Languages Ladder and Asset Languages. Lid King, Director of the National Languages Strategy, provided background to the Languages Ladder. Kate Green of the Department for Education and Skills (DfES) introduced the Languages Ladder and Asset Languages. Neil Jones presented on 'Setting standards within a multilingual proficiency framework: Asset Languages'. As made clear by the methodology being developed for linking exams to the CEFR, the validity of any claim to alignment begins with demonstrating the validity of an exam for its particular purpose. Neil discussed the purposes which the Asset Languages assessment framework sets out to serve, arguing that a proficiency framework like the CEFR is just as much a framework for learning, and that constructs of proficiency should take clearer account of the specific context of groups of learners compared within the framework. This is of particular importance for Asset Languages, with its remit to accredit useful language skills across a wide range of languages, levels and contexts of learning.

Other sessions focused on the candidate's perspective (Kristina Hedges); ensuring positive washback with Welsh exams (Emyr Davies) and the European Language Portfolio (ELP). Joanna Panthier provided background on the Council of Europe and introduced the Council of Europe's ELP while David Little discussed case studies of how the portfolio is being used to benefit learners in Ireland.

## Language Testing Forum 2005

The Language Testing Forum took place at Downing College, Cambridge on 25–27 November 2005. There were pre-conference events, including a workshop led by Ardeshir Geranpayeh and Andrew Somers (Cambridge ESOL) on 'Using Structural Equation Modelling in Construct Investigation'. In the afternoon, staff from the Research and Validation Group demonstrated the Test Construction Cycle at Cambridge ESOL. Other contributions by Cambridge ESOL staff to this event included papers by Ardeshir

Geranpayeh who spoke about 'Language Proficiency Revisited: Demystifying the CAE Construct' and David Thighe who presented 'The International Legal English Certificate: Issues with Developing a Test of English for Specific Purposes'. Ardeshir reported on empirical research investigating the underlying constructs of CAE examinations using Structural Equation Modelling (SEM). In this model, elements of communicative language ability, overall proficiency and its divisibility to language skills are married up to form a communicative model of language proficiency. That is, while there exists an overall communicative language ability, such ability is divisible by skills and language elements. To examine whether the empirical evidence supports the assumptions made above, several plausible models for the CAE examinations were constructed. The viability of each model was tested using SEM techniques. The best fit indices came from a Correlated-Trait (CT) model. It suggested that the Cambridge model of language proficiency for CAE is based on a componential aspect of communicative language ability whereby each component assesses a very different aspect of language proficiency. David's paper described the trial of the International Legal English Certificate (ILEC), reporting on the findings and engaged with such questions as, 'How specific to the legal environment should the test content and test language be?', and 'Will the examination necessarily test candidates' legal knowledge as well as their language ability?'

In a paper entitled 'The Uneasy Guest at the Examiner's Table? Incorporating the Test-Taker in Test Development', Tony Green (Cambridge ESOL), and Barry O'Sullivan (University of Roehampton) argued for the importance of incorporating a clear definition of the test-taker in test development or revision projects. They suggested that failure to fully describe and account for the relevant features of the population for which a test is intended might result in tests that are either biased towards or against particular groups or individuals. However, the literature relating to the test-taker is surprisingly limited. O'Sullivan (2000) has provided a framework for describing test-takers including *experiential, psychological* and *experiential* characteristics. Tony and Barry used this framework to discuss how characteristics under each of the headings may be accessed and accounted for in Speaking test design and delivery. They outlined the challenges raised for test providers and suggested avenues for future research activity.

An open discussion on 'Good Language Testing Practice' on the Saturday afternoon was led by Alan Davies (University of Edinburgh), Liz Hamp-Lyons (Hong Kong University), Nick Saville (Cambridge ESOL), Cyril Weir (University of Luton) and chaired by Lynda Taylor (Cambridge ESOL). Around 30 people participated in the Forum and most people presented a paper, workshop or poster at the event. Look out for the next LTF happening in Reading in November 2006.

**References**


Davies, A (1991) *The native speaker in applied linguistics,* Edinburgh: Edinburgh University Press.

Medgyes, P (1994) *The non-native teacher,* London: Macmillan.

Weir, C (2005) *Language Testing and Validation: An evidence-based approach,* Oxford: Palgrave.


# Recent publications of interest

The latest issue of the international journal *Language Assessment Quarterly* (Vol 2, No 4) contains a fascinating interview between Nick Saville and Prof John Trim. Although he would not consider himself a language tester, John Trim has followed the trends in language assessment since the 1960s and his own work, particularly as a co-author of the Waystage and Threshold levels and most recently the Common European Framework of Reference (CEFR), has been very influential in language testing circles in recent years. The interview explores the background and context in which the CEFR was produced and talks about the formative influences that guided the thinking which informed this work. The same journal issue contains a positive review by Dr Annie Brown of Volume 14 in our Studies in Language Testing series – *A Qualitative Approach to the Validation of Oral Language Tests* by Prof Anne Lazaraton, now at the University of Minnesota.

The topic of globalisation and the English language is increasingly discussed at language teaching/testing conferences as well as in the literature. Recent decades have seen a growth in the description of English varieties used around the world raising interesting issues for teachers and testers in terms of which linguistic models should be adopted for pedagogic and assessment purposes. Some of these issues are debated in a pair of articles in the Point/Counterpoint section of the latest issue of the *ELT Journal*, published by Oxford University Press (60/1, Jan 2006). In an article entitled 'The spread of English as an international language: a testing time for testers', Jennifer Jenkins (Kings College, London) presents a view from applied linguistics; she argues that recent changes in both users and uses of English have become so far-reaching that a substantial overhaul of English language testing is required on the grounds that teachers and learners alike will be reluctant to embrace any curriculum change that is not reflected in the targets set by the major examination boards. In a response article – 'The changing landscape of English: implications for language assessment' – Lynda Taylor (Cambridge ESOL) presents the view from the language testers' perspective; she discusses the key factors which frame how examination boards, particularly Cambridge ESOL, deal with English varieties and discusses the contribution that the language testing community can make to increase our understanding of language variation.