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 24 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 frameworks in assessment and their impact on language tests, teaching awards and the various stakeholder groups who take our tests or make decisions based on them. A key framework is the *Common European Framework of Reference* (CEFR) which has a growing influence on language testing organisations and stakeholders worldwide. We reflect in this issue the provision we make for testing languages other than English (Asset Languages) and how we test English in other domains such as Legal, Academic and Business English.

In the opening article Lynda Taylor and Neil Jones discuss the relationship of Cambridge ESOL's exams with the Council of Europe's CEFR along four perspectives: historical, conceptual, empirical and evolutionary. Next David Thighe describes how a new test, the International Legal English Certificate (ILEC) is being related to the CEFR using the three stage process suggested by the Council of Europe's *Pilot Manual*, a working document that outlines how exams can be aligned to the CEFR. These articles are followed by a Research and Development update for Asset Languages.

The following two articles describe how Asset Languages is being linked to the CEFR. Tamsin Walker considers how learners taking Asset exams can be said to be linked to the CEFR, describing learner-based standard-setting and suggesting a holistic approach to assessment. Karen Ashton reports on the development of a Can Do self-assessment tool for learners of German, Japanese and Urdu which aims to ensure that the difficulty of tasks and ability of learners taking tests in different languages are comparable.

Next Chris Hubbard, Susan Gilbert and John Pidcock report on a Verbal Protocol Analysis (VPA) study into how CAE Speaking test raters make assessments in real time. They consider the appropriacy of a VPA methodology and how raters use a framework of assessment criteria (a rating scale). Stuart Shaw then considers rating scales for Writing, in his concluding article on the IELTS Writing Revision Project. He focuses on the qualitative analysis of a global survey on the revised IELTS Writing rating scale. Both raters and administrators were surveyed; the latter being a key stakeholder group rarely foregrounded in research studies.

Nadežda Novaković then describes the first year of the Teaching Knowledge Test (TKT) in terms of the candidates' profile and their performance. She explains how Cambridge ESOL is measuring TKT candidates' language proficiency to determine if this affects their performance on the TKT.

We end this issue with the call for applications for the twelfth round of research funding under the IELTS Joint-funded Research Program.

Cambridge ESOL exams and the Common European Framework of Reference (CEFR)

LYNDA TAYLOR AND NEIL JONES, RESEARCH AND VALIDATION GROUP

Introduction

A previous Research Notes article explored issues of test comparability and the role of comparative frameworks as communicative tools (Taylor 2004). One framework which has a growing role for language testers is the Common European Framework of Reference (CEFR; Council of Europe 2001). At Cambridge we are often asked about the relationship between our ESOL exams and the CEFR; the nature of this relationship can be considered from four complementary, sometimes overlapping, perspectives.[1]

The historical perspective

The origins of the CEFR date back to the early 1970s when the Council of Europe sponsored work within its Modern Languages Project to develop the Waystage and Threshold levels as sets of specified learning objectives for language teaching purposes. These two levels were designed to reflect achievable and meaningful levels of language competence, at a relatively low proficiency level, and to form part of a European unit/credit system for adult language learning. They defined levels of functional competence among language users forming the basis for curriculum, syllabus, and later assessment design.

In the late 1980s Cambridge was one of several stakeholder organisations (with the British Council and BBC English) to provide funding and professional support for revising Threshold and Waystage (Van Ek and Trim 1998a, 1998b); the revised level descriptions underpinned test specifications for a revised PET exam in the mid 1980s and a new KET exam in the early 1990s.

Linguistic and functional description of a third, higher proficiency level began in the 1990s, with support and participation on this occasion from the Association of Language Testers in Europe (ALTE); work on this third level took account of FCE and led to the publication of Vantage in 1999 (Van Ek and Trim 2001). As work extended on level descriptions for English, so the concept of a framework of reference levels began to emerge and to take on a more concrete form.

The conceptual perspective

In part, emergence of a framework formalised conceptual levels with which ELT learners, teachers and publishers had operated for some years – with familiar labels such as 'intermediate' or 'advanced'. Dr Brian North, one of the CEFR's authors, confirms its origins in traditional English Language Teaching levels:

> *The CEFR levels did not suddenly appear from nowhere. They have emerged in a gradual, collective recognition of what the late Peter Hargreaves (Cambridge ESOL) described during the 1991 Rüschlikon Symposium as "natural levels" in the sense of useful curriculum and examination levels.*
>
> *The process of defining these levels started in 1913 with the Cambridge Proficiency exam (CPE) that defines a practical mastery of the language as a non-native speaker. This level has become C2. Just before the last war, Cambridge introduced the First Certificate (FCE) – still widely seen as the first level of proficiency of interest for office work, now associated with B2. In the 1970s the Council of Europe defined a lower level called "The Threshold Level" (now B1), originally to specify what kind of language an immigrant or visitor needed to operate effectively in society. Threshold was quickly followed by "Waystage" (now A2), a staging point half way to Threshold. The first time all these concepts were described as a possible set of "Council of Europe levels" was in a presentation by David Wilkins (author of "The Functional Approach") at the 1977 Ludwighaven Symposium…(North 2006:8).*

Cambridge's upper-intermediate level CAE exam, introduced in 1991, helped bridge the gap between FCE and CPE and was proposed as C1. Lastly, a lower Breakthrough level was proposed as A1. These six levels (A1-C2) thus constituted a 'language ladder', providing a pathway for upward progression in language teaching and learning with explicit opportunities to evaluate and accredit learning outcomes along the way. The Cambridge Main Suite exams (KET, PET, FCE, CAE and CPE) were already providing well-established and recognised accreditation 'stepping stones' along this pathway.

Emergence of these common reference levels, with their contributory elements such as language courses, public examinations, and published coursebooks, was formally confirmed through the Common European Framework project; managed between 1993 and 1996 by the Council of Europe with significant input from the Eurocentres organisation, the overarching aim was to construct a common framework in the European context which would be transparent and coherent, to assist a variety of users in defining language learning, teaching and assessment objectives. A major strength was that it would build upon the shared understanding which already existed among teachers and other ELT stakeholders in the European context, but would also resolve some difficulties of relating language courses and assessments to one another; it would provide a common meta-language to talk about learning objectives and language levels and encourage practitioners to reflect on and share their practice. It's worth remembering that this took place in a larger context where notions

[1] This article is based on a presentation given at IATEFL Harrogate in April 2006 and we are grateful to Dr Brian North for his helpful comments on an early draft.

of a socio-political and economic community in Europe were rapidly taking shape; an early motivation for revising Waystage and Threshold in the late 1980s had been their relevance to educational programmes of language learning for European citizenship.

Notions of framework development linked to language learning progression were nothing new. Wilkins' 1977 set of levels has already been referred to. In the UK context, the English Speaking Union (ESU) set up its 'framework project' in 1985 to devise a comprehensive frame of description for comparing the various examinations of the main English language boards (Taylor 2004). In the wider context of Europe, ALTE members were also by the early 1990s working systematically to co-locate their qualifications across different European languages and proficiency levels within a shared framework of reference. The aim was to develop a framework to establish common levels of proficiency in order to promote the transnational recognition of certification in Europe. The process of placing ALTE members' exams on the framework was based on content analysis of the tests, the creation of guidelines for the quality production of exams, and the development of empirically validated performance indicators or Can Do statements in different European languages (see ALTE website www.alte.org). The resulting five-level ALTE Framework developed simultaneously during the mid-1990s alongside the six-level CEFR published in 1997. Since the two frameworks shared a common conceptual origin, similar aims – transparency and coherence – and comparable scales of empirically developed descriptors, Cambridge ESOL and its ALTE partners decided to conduct several studies to verify their alignment. This was achieved mainly through the ALTE Can Do Project in 1998-2000 (see below). Following publication of the CEFR in 2001 the ALTE members adopted the six CEFR levels (A1-C2).

One of the strengths of this conceptual approach to framework development has undoubtedly been its 'organic' development. Even in 1991, qualifications existed for other languages that could also be confidently associated with what were to become the CEFR and ALTE levels, including: the new advanced level DALF (Diplôme Approfondi de Language Française) at C1; the Zertifikat Deutsch (ZD) at Threshold (B1); and the Kleines Deutsches Sprachdiplom (KDS) commonly considered an equivalent to Cambridge's CPE (C2).

The empirical perspective

Shared understanding among teachers, publishers and language testers enabled the framework concept to function quite well without extensive underpinning from measurement theory and statistics; but measurement theory has become increasingly important as attempts have been made to validate aspects of the CEFR empirically (North and Schneider 1998, North 2000a) and to link assessments to it (North 2006b).

Syllabus designers, coursebook publishers and language test providers worldwide, including Cambridge ESOL, seek to align their exams to the CEFR for reasons of transparency and coherence; claims of alignment can also assist in marketing communications to try and gain a competitive edge. However, any claim of alignment needs to be examined carefully; simply to assert that a test is aligned with a particular CEFR level does not necessarily make it so, even if that assertion is based on an intuitive or reasoned subjective judgement. To some extent, alignment can be achieved historically and conceptually as we have seen, but empirical alignment requires more rigorous analytical approaches; appropriate evidence needs to be accumulated and evaluated.

The ALTE Can Do Project (Jones 2001, 2002) was one of the empirical approaches used by Cambridge ESOL for aligning its original five levels with the six-level CEFR. Other empirical support for alignment comes from Cambridge's item-banking methodology underpinning our approach to all test development and validation (Weir and Milanovic 2003). The Cambridge-TOEFL Comparability Study, conducted in 1987-90 (Bachman et al 1995) highlighted how far the UK-based assessment tradition had relatively underplayed the psychometric dimension; for Cambridge ESOL this established an empirical imperative and we invested heavily in approaches and systems to address measurement issues such as test reliability and version comparability. Latent trait methods have been used since the early 1990s to link the various Cambridge levels onto a common measurement scale using a range of quantitative approaches, e.g. IRT Rasch-based methodology, alongside qualitative research methods.

More recently, Cambridge ESOL has supported the authoring and piloting of the Council of Europe's Manual Relating Language Examinations to the CEFR (Figueras et al 2005) which presents a linking process based on three sets of procedures:

Specification of the content and purpose of an examination

Similar procedures were conducted when the PET and KET test specifications were originally based upon Threshold and Waystage levels, and the ALTE partners' exams were aligned within the ALTE Framework; an extensive range of documentation for all our exams (test specifications, item writer guidelines, examiner training materials, test handbooks and examination reports) assists in specifying the content and purpose of existing and new exams with direct reference to the CEFR.

Standardisation of interpretation of CEFR levels

Suitable standardised materials are needed for assessment personnel and others to benchmark their tests against CEFR levels. Cambridge has helped develop such materials by supplying calibrated test items and tasks from our Main Suite Reading and Listening test item banks together with exemplar Speaking and Writing test performances from our writing examiner coordination packs and Oral Examiner standardisation materials at each CEFR level; a set of benchmarking materials, incorporating both classroom-based and test-based materials, is now available from the Council of Europe on CD or DVD.

Empirical validation studies

Empirical validation studies are a greater challenge sometimes requiring specialist expertise and resources; Cambridge ESOL is among a relatively small number of examination providers undertaking this sort of research, partly through our routine item-banking and test calibration methodology and also through instrumental research and case studies such as the Common Scale for Writing Project (Hawkey and Barker 2004).

The evolutionary perspective

The CEFR remains 'work in progress'; it will continue to evolve as experience grows among those who use it in various ways and contexts, and as they reflect on that use. For many it already provides a useful frame of reference, offering practical guidance for their thinking and doing. Others have expressed concerns about its application: within the language testing community some fear use of the CEFR as an instrument for 'harmonisation' of policy/practice (Fulcher 2004); others question how far the CEFR provides a suitable instrument for operational test development (Weir 2005). In response, the CEFR authors emphasise the original intention of the Framework as a means of valuing and encouraging diversity, and remind us that the CEFR is not a 'cookbook' or 'how to' document. Perhaps the real value of the CEFR lies in it being used as a heuristic rather than prescriptively; it needs to be interpreted thoughtfully and intelligently if it is to be meaningful and have local validity.

Another useful role for the Framework in assessment could be in matters of quality assurance, not just to improve systems and procedures but to support the growing professionalisation of personnel and institutions involved in language learning, teaching and assessment. North (2006) notes that the scheme outlined in the Manual 'reflects the three step process of any Quality Management system (Design, Implementation, Evaluation)'. This view echoes Cambridge ESOL's long-standing commitment to addressing quality assurance issues. In the early 1990s ALTE produced its professional Code of Practice and has since then elaborated the concept of quality assurance in language testing by developing quality management instruments. Like the CEFR, the ALTE Code of Practice offers the practitioner community a common frame of reference and a shared meta-language for reflecting on and evaluating policy and practice – ensuring the door is always open for improvement.

Since 2001, the CEFR has also been a source of inspiration or a catalyst for other initiatives; one is the innovative European Language Portfolio (ELP) developed to support the language learning and teaching community with input from the EAQUALS organisation and the ALTE partners; another is the recently launched English Profile Project to develop a comprehensive set of Reference Level Descriptions for English using the CEFR levels as a springboard.

Conclusion

Today the CEFR plays a key role in language and education policy within Europe and the wider world – perhaps in ways not originally envisaged by its authors. Within Europe it is believed to serve policy goals of fostering linguistic diversity, transparency of qualifications, mobility of labour, and lifelong language learning. Beyond Europe it is being adopted to help define language proficiency levels with resulting implications for local pedagogy and assessment. For Cambridge ESOL it offers a valuable frame of reference for our work and for our stakeholder community; as it evolves, we look forward to continuing to make an appropriate professional contribution to its development.

Could it be argued that Cambridge ESOL exams 'embody' the Common European Framework? That will be for others to judge based on evidence presented here and elsewhere. It partly depends on how the word 'embody' is defined; but there does exist a growing body of evidence to support a claim that Cambridge exams contain and express the CEFR as an important feature, that they include the CEFR as part of their structure, and that they express or represent the CEFR in a variety of ways. Such embodiment is a natural outcome of several factors, such as historical legacy, conceptual synergy, and empirical underpinning. Extending the biological metaphor, we could envisage how the relationship between the CEFR and Cambridge ESOL exams will continue to evolve, partly due to the genetic makeup of the relationship itself and also as a result of external environmental factors in a changing world.

To celebrate his 80th birthday in 2004, Professor John Trim, one of the authors of the CEFR, was interviewed for Language Assessment Quarterly. In the interview, he describes the aspirations behind the Framework: 'What we were aiming at was something which will be a common reference point that people working in different fields and people using it for entirely different things and in very different ways could refer to in order to feel that they were part of a common universe' (Saville 2005:281). This focus on individual practitioners as the agents of activity is a welcome reminder that it is people, rather than frameworks, systems, or procedures, who are – or who should be – at the heart of what happens in language learning, teaching and assessment, i.e. learners, teachers, teacher trainers, course and syllabus designers, textbook writers, language test providers – anyone who is a stakeholder in the ELT or ESOL constituency, or who is a member of another language learning community.

Ultimately, it may be unhelpful to talk about 'embodiment' in relation to a course syllabus or an assessment tool; of greater interest and importance, both to the developers of the CEFR and to Cambridge ESOL, are surely the populations of human beings directly involved in language learning, teaching and test-taking, whether at the group or the individual level. The quality of the relationship between the CEFR and Cambridge ESOL exams is perhaps best judged by the extent to which together they enable language learning to flourish, encourage achievements to be recognised and so enrich the lives of individuals and communities.

References and further reading

Bachman, L F, Davidson, F, Ryan, K and Choi, I C (1995) *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*, Cambridge: UCLES/Cambridge University Press.

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

Figueras, N, North, B, Takala, S, Verhelst, N and Van Avermaet, P (2005) Relating Examinations to the Common European Framework: a Manual, *Language Testing* 22 (3), 1–19.

Fulcher, G (2004) Deluded by artifices? The Common European Framework and harmonization, *Language Assessment Quarterly*, 1 (4), 253–266.

Hawkey, R & Barker, F (2004) Developing a common scale for the assessment of writing, *Assessing Writing* 9/2, 122–159.

Jones, N (2001) The ALTE Can Do Project and the role of measurement in constructing a proficiency framework, *Research Notes* 5, 5–8.

—(2002) Relating the ALTE Framework to the Common European Framework of Reference, in Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*, Strasbourg: Council of Europe Publishing, 167–183.

North, B (1995) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement, *System* 23, 445–465.

—(2000a) *The development of a common framework scale of language proficiency*, New York: Peter Lang.

—(2000b) Linking Language Assessments: an example in a low-stakes context, *System* 28, 555–577.

—(2006) The Common European Framework of Reference: Development, Theoretical and Practical Issues, paper presented at the symposium 'A New Direction in Foreign Language Education: The Potential of the Common European Framework of Reference for Languages', Osaka University of Foreign Studies, Japan, March 2006.

North, B and Schneider, G (1998) Scaling Descriptors for Language Proficiency Scales, *Language Testing* 15 (2), 217–262.

Saville, N (2005) An interview with John Trim at 80, *Language Assessment Quarterly* 2 (4), 263–288.

Taylor, L (2004) Issues of test comparability, *Research Notes* 15, 2–5.

Van Ek, J A and Trim, J L M (1998a) *Threshold 1990*, Cambridge: Cambridge University Press.

—(1998b) *Waystage 1990*, Cambridge: Cambridge University Press.

—(2001) *Vantage*, Cambridge: Cambridge University Press.

Weir, C J (2005) Limitations of the Common European Framework for developing comparable examinations and tests, *Language Testing* 22 (3), 281–300.

Weir, C J and Milanovic, M (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002*, Studies in Language Testing Vol. 15, Cambridge: UCLES/Cambridge University Press.

Placing the International Legal English Certificate on the CEFR

DAVID THIGHE, RESEARCH AND VALIDATION GROUP

Introduction

The International Legal English Certificate (ILEC) is a test of English in a legal environment designed by Cambridge ESOL in partnership with Translegal, a leading firm of lawyer-linguists based in Europe. It is aimed at people working in or studying law who require evidence of their proficiency in English in a legal working environment. ILEC comprises four components (Reading, Writing, Listening and Speaking) and assesses candidates' language proficiency by providing texts and tasks that candidates may be expected to meet in their working environment. Translegal have provided much assistance in suggesting suitable texts for use in the examination and in task design. A period of extensive trialling of test versions was completed in July 2005 and the first administration of the examination is in May 2006. ILEC is aimed at candidates who are at B2 or C1 level on the *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2001) and the examination reports passing grades as C1 and B2.

The placing of ILEC on the CEFR allows candidates and teachers to ascertain what general level of English language proficiency is required to be successful in the examination. It also allows comparisons of candidates' achievement in ILEC with that of candidates in other examinations placed on the Framework and provides a description of what successful ILEC candidates can do in English in a legal working environment. Placing ILEC on the CEFR then is of practical use to a number of stakeholders. This article examines the applicability of doing this and argues that such a procedure is meaningful. It also attempts to show how in practice this can be done and how evidence of the validity of this process can be established.

Placing ILEC on the CEFR

The *Preliminary Pilot Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (henceforth Pilot Manual*; Council of Europe 2003) provides a systematic and comprehensive methodology for how to place an examination on the CEFR. It details a three stage process (following an initial familiarisation stage) of:

1. Specification of examination content – this stage makes a claim that a test relates to the CEFR from a study of the tasks and texts in the test.
2. Standardisation of judgement – this stage substantiates the claim by standardising markers' judgements to the CEFR.
3. Empirical validation though analysis of test data – this stage establishes the claim by providing evidence independent of the judgement of item writers and examiners.

For ILEC the systematic approach described above was adopted to show that it tests at levels B2 and C1 on the CEFR and to highlight research needed to further establish this claim.

Specification of examination content

Before a test can be related to the CEFR it must be shown that the test is valid in that it tests what it claims to test, and is *reliable* in that it can accurately and consistently measure candidates' ability. Without these any claim to being on an external framework such as the CEFR is redundant. For ILEC, test designers and item writers have worked closely with Translegal to ensure that tasks and texts reflect the type of activities that lawyers and law students may be expected to do in their legal working environment. The success of this collaboration with Translegal is reflected in the high face validity that all ILEC components achieved when trialled on the target test population. For example, for the Reading component 90% of all candidates who expressed an opinion agreed or strongly agreed that the topics, texts and language were authentic, with only 10% disagreeing. Similar results were found for the other components.

With a population similar in range of ability to that of other Cambridge ESOL examinations such as the First Certificate in English (FCE) and the Business English Certificates (BEC), similar measures of Reliability were found, indicating that the test contains an appropriate number and quality of items and tasks to measure candidates' performance consistently. For example, in a trial of the ILEC Reading component on 88 candidates, Cronbach's alpha (a measure of the internal consistency estimate of Reliability for a test) was 0.89, this compares to 0.84 for one session of a BEC Vantage examination based on several thousand candidates.

The first stage in the process of relating an exam to CEFR given in the *Pilot Manual* recommends a detailed specifications document is written for a test to form the basis for the provision of evidence of that test's validity. For ILEC a test specifications document has been completed based on a socio-cognitive model for test validation provided by Weir (2005a). These specifications allow the test designers to examine how every decision about the content and scoring methods to be used in the test will affect how candidates think and act during the test. Primarily the specifications allow us to consider whether candidates think and act in the same way during the test as they would in the target language use situation, thus showing the interactional authenticity of the test.

Stage 1 is not only about providing evidence of the validity of the test in its own right (what is described as *internal validity* in the *Pilot Manual*), it also recommends that a process of *external validity* is undertaken where test designers review the tasks and functions to see if they can be matched up to descriptors of candidate performance at the CEFR levels (see Council of Europe 2001). ILEC test designers and item writers have included tasks in the test which cover the contexts of use, functions and tasks at B2 and C1 levels as provided in the *CEFR*. As all item writers for ILEC have experience of working on established Cambridge ESOL examinations such as FCE, they bring to ILEC a working knowledge of CEFR levels. An additional tool to help item writers and designers is Item Response Theory (Rasch one-parameter model) which provides an indication of the level of a task on a common scale of task difficulty. The CEFR level of a task can be found through the trialling of that task together with other tasks which have already been calibrated to CEFR levels on a representative sample of the target candidature.

Standardisation of judgement

The processes of internal and external validity allow us to claim that ILEC tests at B2 and C1 on the CEFR. Stage 2 (Standardisation of judgement) focuses on how to ensure that item writers and examiners are adequately trained to make judgements relating ILEC to the CEFR over an extended period of test administration.

For the productive skills (Writing and Speaking) ILEC examiners, as in established Cambridge ESOL examinations, take part in a process of Recruitment, Induction, Training, Co-ordination, Monitoring and Evaluation (RITCME; see Mitchell Crow and Hubbard 2006). For example Oral Examiners must be experienced language teachers with a demonstrated level of language competence. Prior to training, successful applicants are sent a self-access Induction Pack, which contains a video, worksheets, Instructions to Oral Examiners, and provides applicants with the main principles of the Cambridge ESOL Speaking tests. Following induction, applicants are invited to a training session where they are given training and practice in the roles of interlocutor and assessor. These skills include both generic skills which apply to the conduct of all the speaking tests, and skills specific to particular examinations such as ILEC. To ensure that all practising Oral Examiners are marking to acceptable standards, all examiners must attend a co-ordination meeting on an annual basis. Newly trained examiners must attend a co-ordination session before their first examining session and then on an annual basis. Once every two years all Cambridge ESOL Oral Examiners are monitored by a Team Leader. The Team Leader monitors an Oral Examiner's performance as interlocutor and assessor by monitoring at least two live interviews.

For the standardisation of judgement in relation to the CEFR in ILEC productive skills components it is necessary for the RITCME process described above to be linked to the CEFR levels. It is planned that experts trained in the assessment of CEFR levels examine a range of live writing scripts and video-ed live oral performances and assess these in terms of CEFR levels. These results will then be compared to the results achieved in the live administration of the test. It will be necessary for these experts to follow an initial familiarisation process as suggested in the *Pilot Manual* such as discussing aspects of the CEFR, sorting scales of descriptors and using self-assessment instruments. It is also advocated that this cadre of CEFR trained experts assess the oral performance videos and writing scripts used in the training process described above.

ILEC's receptive skills components (Reading and Listening) comprise a combination of multiple-choice items and items requiring short responses marked to a pretested answer key. Both types of items are dichotomously marked (correct or incorrect). Standardisation of judgement with these components requires focus on the process of item writing. It is proposed that experts, familiar with the CEFR and its functional/situational descriptors, examine tasks in these components to assess whether a correct response indicates a CEFR level of B2 or C1. Item Response Theory applied to the first live administrations will also provide evidence of the relation of the receptive components to the CEFR and will input into standard-setting of the cut-off points at levels B2 and C1 for the whole examination.

Empirical validation through the analysis of test data

Stage 3 of the *Pilot Manual* represents an attempt to provide empirical evidence for the claim made in Stage 1 and substantiated in Stage 2 that the test can indicate a candidate's level on the CEFR.

For ILEC it is proposed that the cadre of CEFR trained experts together with professionals with experience of teaching and working in a legal environment review existing CEFR functional descriptors to produce a set of short Can Do statements that are legally oriented in that the situations they describe are those that may be met in a legal working environment. The ALTE Can Do statements, which have been mapped onto the CEFR (Jones and Hirtzel 2001), will be used as a starting point for this work. These legal Can Do statements will represent a range of CEFR levels A2 to C2 but will contain predominantly what are expected to be B2 and C1 statements. These will be presented to candidates in the first administrations of the live test and candidates will be asked to indicate which Can Do statements they think apply to themselves. Item Response Theory will be used to indicate those statements that do not fit in a hierarchy of candidate ability or Can Do difficulty. Those legally oriented statements that do not fit or discriminate will be rejected from the set of statements. The remaining statements will be compared to the performance of candidates in the test to see if the CEFR level of candidates achieved in the test and its components match the CEFR levels of the statements. This process will provide evidence of the claim that ILEC is at CEFR levels B2 and C1 and also provide a set of legally oriented Can Do statements that can be reported on the statement of results for candidates. These may provide stakeholders with a description of what successful ILEC candidates can be expected to be able to do in English in a legal working environment.

Conclusion

This article has described our completed and planned research to validate the claim that successful candidates in ILEC can be placed meaningfully on the CEFR. It uses the three stage approach of Specification of examination content, Standardisation of judgement and Empirical validation advocated in the *Pilot Manual*.

Recent evaluations of the CEFR by Weir (2005b) and Hardcastle (2004) criticise the framework for not providing context-specific statements of what candidates can do at CEFR levels and for assuming that functional statements are in the same hierarchy of acquisition across different languages and language domains. The research advocated in this article, of administering legally oriented Can Do statements to candidates for self assessment, will go some way to verify that context-specific functional Can Do statements meaningfully apply to users of English in a legal working environment and allow us to align ILEC on the CEFR.

References and further reading

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

—(2003) *Preliminary Pilot Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Strasbourg: Language Policy Division.

Hardcastle, P (2004) Test Equivalence and Construct Compatibility across Languages, *Research Notes* 17, 6–11.

Jones, N and Hirtzel, M (2001) Appendix D: The ALTE Can-do Statements, in Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Mitchell Crow, C and Hubbard, C (2006) ESOL Professional Support Network Extranet, *Research Notes* 23, 6–7.

Taylor, L (2004) Issues of Test Comparability, *Research Notes* 15, 2–5.

Weir, C (2005a) *Language Testing and Validation: An Evidence-based Approach*, Hampshire: Palgrave Macmillan.

—(2005b) Limitations of the Common European Framework of Reference (CEFR) in Developing Comparable Examinations and Tests, *Language Testing* 22 (3), 281–300.

Asset Languages Research and Development

At present Asset Languages has assessments available in eight languages – French, German, Spanish, Italian, Panjabi, Urdu, Japanese and Chinese. Assessments in a further 11 languages are currently being developed. Ensuring that the standard required in Asset Languages assessments is communicated and comparable across these languages is a challenge that the team faces. In order to meet this challenge, exemplar videos of Asset Languages speaking tests were developed in English in December 2005. These videos were piloted in training sessions, both in initial standardisation meetings and in cross-language standardisation meetings. These videos, used in conjunction with CEFR descriptors, have proved extremely useful in communicating the level to external examiners and moderators and in arriving at good agreement of the level both for a particular language and across languages. Similar exemplar materials are now being developed for writing and these will be piloted in training sessions in 2006.

In addition to ensuring comparability is maintained across languages, the area of comparability across frameworks is a major focus of the Asset Languages research agenda. Several projects have been initiated investigating the relationship between Asset Languages levels and the levels of current qualifications within the UK education system. The results of this research should enable the Asset Languages team to give additional information and guidance to teachers in the UK. For more information on Asset Languages visit www.assetlanguages.org.uk

Linking learners to the CEFR for Asset Languages

TAMSIN WALKER, RESEARCH AND VALIDATION GROUP

Introduction

This article considers the proposition that 'It is learners, rather than tests, that are linked to the CEFR' and shows how we might justify the claim that Asset Languages locates candidates at a given CEFR level. The kind of evidence and reasoning that is required beyond simply describing features of the tests themselves is discussed.

Asset Languages is a suite of exams currently implemented in England, which will eventually assess 26 languages at all levels of proficiency across primary, secondary and adult contexts (Jones, Ashton and Chen 2005). The *Common European Framework of Reference for Languages* (CEFR, Council of Europe 2001), in particular its tables of descriptors scaled to the six reference levels A1 to C2, is used to align different language qualifications into a single framework. The Council of Europe has produced a *Pilot Manual* (Council of Europe 2003) which describes in detail how this can be done. The manual concentrates on defining the level of an exam by focusing on the test task. An alternative approach is learner-based standard-setting: data is gathered on what the candidates themselves are able to do, either through rating of candidates' performance, teacher evaluations, or self-rating.

Relating Asset performance tests to the CEFR

The skills of writing and speaking are assessed for Asset as performance tests, that is tasks are chosen to elicit a sample of written text or speech, including spoken interaction, which aims to reflect real life tasks, such as asking and answering factual questions, giving a presentation, or writing a short postcard. These tests are assessed according to two criteria: 'communication' assesses the candidates' communicative ability, including fulfilment of the task; 'language' assesses the linguistic competence shown by the candidate.

In terms of relating these tests to the CEFR descriptors, there is a demonstrable link to some descriptors, for example, 'Can give a prepared straightforward presentation' (Addressing Audiences, B2, Council of Europe 2001:60). Performance tasks still present a number of issues to be resolved in demonstrating a link to the CEFR, however. Only a subset of descriptors are covered, since only a sample of performance can be elicited for practical reasons. McNamara (1996:51) states 'The practical problem in performance assessment is the generalization from one observable instance of behaviour to other unobservable instances'. How a candidate might fulfil descriptors related to interaction in the travel agents, for instance, can only be judged by extrapolating the performance. Another issue is the number of variables which can affect a candidate's performance (Milanovic and Saville 1996:8), such as interlocutor behaviour, candidate age, test conditions and preparation, all relevant to Asset Languages, with its range of learning contexts. A consistent approach to these factors needs to be taken, which can be achieved most effectively by the use of exemplars. This applies equally to the interpretation of the CEFR descriptors: what do 'simple' 'basic' or 'varied' mean, and what constitutes fulfilling a statement successfully?

The main basis for a claim of alignment to a CEFR level for the Asset performance tests is through qualitative expert judgement. To validate these judgements, these claims need to be corroborated by empirical evidence. The following three stages can be used to provide a reasoned argument.

Standard-setting using expert judgement

This stage provides a claim for Asset tests in terms of CEFR levels based on qualitative judgement; its outcome is a set of graded exemplars which can be used to train examiners and moderators, and ensure consistency of grading. English exemplars are particularly useful in providing a central point of reference which can be used to ensure cross-language comparability, as well as comparability to Cambridge ESOL levels. Not only have the Cambridge ESOL levels (which have come about through refinement of tests over many years) been aligned with the CEFR through empirical evidence, but according to one of the authors of the CEFR, this is itself based to some degree on the Cambridge level system (North 2004). Rating of different languages by multi-lingual examiners and moderators also helps to ensure a consistent standard.

To rate samples in terms of CEFR levels, familiarisation with the salient features of the Common Reference Levels (Council of Europe 2003:19), examples of which include 'Can enter unprepared into conversation on familiar topics' (B2) or 'Can order a meal' (A2), is required.

Establish a quantitative link to the CEFR using Can Do descriptors

This phase provides learner-centred data which complement the rater judgement in the previous phase. Can Do descriptors relating to Asset tasks and contexts should be compiled, as described in Karen Ashton's article in this issue. The inclusion of CEFR descriptors as anchor tasks means that a single calibrated scale can be obtained, which can be related to the CEFR levels. Asset exam results can be compared to ability as measured by the Can Do scale, to establish whether the Asset assessments are discriminating between candidates of differing abilities. To provide evidence that the Asset and CEFR levels align, the cut-off values found for the original CEFR descriptors (Council of Europe 2003:119) can be used. If there is a strong enough correlation for individual results, regression analysis could be used to relate the Can Do scale to the Asset grades, that is, to answer the question 'What ability does a typical Asset candidate of grade x have in Can Do terms?'. If the

CEFR cut-offs agree well with the grading of the Asset candidates, for example candidates achieving Asset Intermediate level have similar abilities to a CEFR B1 level measured in Can Do terms, then this provides quantitative evidence of the claims provided in stage 1.

This stage is particularly useful for identifying differences between groups of learners such as different language learners or age groups. For instance, younger learners might rate a Can Do statement related to making a complaint as more difficult than other groups, because they are unused to using language in this way. Learners of languages which use non-Latin scripts may find basic writing tasks relatively difficult. These findings should be used to inform both item writers and raters, and its effect on rating should be stated clearly in the exam specifications.

Empirical evidence for Asset levels through comparison with other exams

Data on Asset candidates' results or estimated levels on other assessments provide another useful source of corroborating evidence for CEFR level. A standardisation exercise comparing English Asset exemplars with Cambridge ESOL tests would provide information on whether an Asset Intermediate candidate is considered the same level as a PET candidate, for example, and by inference is at level B1 of the CEFR. Similarly, rating of Asset examples alongside CEFR exemplars of speaking and writing will show how Asset levels compare to the CEFR.

In the contexts in which Asset assessments are used, the most widely available data are teachers' National Curriculum (NC) estimates of secondary students. NC levels are described in user-oriented, practical statements, which are roughly equivalent to the lower levels of the CEFR. The relation of NC levels to the CEFR has yet to be determined precisely, but if Asset grades correlate well with NC levels (if one increases as the other increases) this provides additional evidence that Asset is discriminating between different abilities and is at approximately the intended level.

Relating Asset objective tests to the CEFR

The proficiency of candidates on the receptive skills of reading and listening is measured in Asset, as in many assessments, by their scores on tests of short answer or multiple-choice items. A uniform set of task types is used across languages as one basis for a claim of cross-language comparability. There is a less obvious connection between the test construct or the CEFR descriptors, and the responses elicited by these tests. Using the *Pilot Manual's* task-based standard-setting methods using qualitative judgements has proved problematic (Alderson et al 2004).

However, an advantage of objectively marked tests is that the difficulty of the items can be calibrated directly. For each skill, a single scale of item difficulty for all levels can be constructed using candidate response data, providing that trial tests containing items from two levels are administered. Tests can then be constructed to target levels of difficulty. The process for relating objectively-marked tests to the CEFR, although using similar empirical evidence to performance tests, starts with a claim based on quantitative methods of scale construction and standard-setting, rather than qualitative evidence.

Define grade boundaries to align with Cambridge ESOL's model

Defining the boundaries for the six Asset external assessment grades to give a similar progression of ability to that of Cambridge ESOL exams provides an important quantitative claim for the alignment of Asset objective tests to the CEFR (Jones 2005). For initial standard-setting, subjective judgement can be limited to equating the Asset grade boundaries to the ESOL scale at two levels, one high and one low, such as KET and CPE. Replicating the same pattern of proportional gain across different languages gives a quantitative method of achieving cross-language comparability. The grade boundaries will be refined as a result of such empirical findings.

Relate Asset levels to the CEFR using Can Do descriptors

Similarly to the performance tests, a calibrated scale of Can Do descriptors can be used to evaluate test discrimination and to position Asset objective tests in relation to the CEFR. In this case, two ability scales are compared, one from the test results and one from the learner-centred data. The Can Do scale provides a valuable link between the test-oriented language use of the objective tests and real-life language use. As before, findings need to be carefully interpreted: differences between Asset levels and the CEFR (or ESOL) frameworks may be explained by the different learning contexts relevant to the Asset assessments compared to the adult modern foreign language learner context of the CEFR.

Provide empirical evidence through comparison with other exams

Incorporating an anchor task from a CEFR-calibrated reading or listening assessment into an Asset trial test enables both tests to be calibrated on the same proficiency scale, provided the task is carefully chosen to test a similar construct to Asset. If this is the case, a direct comparison of levels can be made. Asset tasks, particularly at the lower levels, could be translated into English to combine with Cambridge ESOL tests, or tasks in European languages could be combined with those from ALTE Partners' exams (see ALTE website for further information www.ALTE.org). In addition, the correlation of teachers' NC estimates with candidates' measured ability levels can be used to evaluate how well the assessments discriminate between different proficiency levels, as well as provide a further indication of the level of Asset grades.

Conclusion

The task of demonstrating a link between Asset Languages assessments and the CEFR is best achieved through comparing and interpreting data from different sources, quantitative and qualitative, task and learner-centred. The test task, and the level of language proficiency it tests, is intrinsic both to the qualitative judgements used in performance tests and the quantitative definition of the objective test grade boundaries. Rather than use the taxonomic view of task difficulty described in the *Pilot Manual*, it seems best to take an holistic approach, and rather than examining each level separately, viewing the progression of stages in the context of the whole framework makes more sense. These approaches are complemented by the learner-centred calibration of Can Do statements and teacher estimates of students levels, giving a wider view of candidates' ability than can be gained from their performance on a particular test. Validation using empirical evidence to compare Asset with other CEFR-related tests provides further evidence to back-up claims made.

The CEFR has been criticised for being poorly defined (Weir 2004). However, the use of the CEFR descriptors has produced a coherent pattern of responses (Jones 2002), indicating an overall agreement in interpretation. The flexibility in the framework allows exams relating to different contexts and languages to be included. What is important is not attempting to conform to a narrow definition of the framework, but to identify, interpret, and clearly specify those areas of the CEFR which do not apply to an assessment, and any areas where the assessment and the CEFR differ.

References and further reading

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S, Tardieu, C (2004) *The Development of Specifications for Item Development and Classification within the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Reading and Listening. The Final Report of the Dutch CEFR Construct Project*, Project Report, Lancaster University, Lancaster, UK, retrieved from: www.eprints.lancs.ac.uk/44/

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

—(2003) *Relating language examinations to the CEFR. Manual; Preliminary Pilot Version*, retrieved from: www.coe.int/T/E/Cultural Co-operation/education/Languages/Language Policy/Manual/default.asp

Jones, N (2002) *Relating the ALTE Framework to the Common European Framework of Reference*, in Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*, Strasbourg: Council of Europe Publishing, 167–183.

—(2005) Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills, *Research Notes* 19, 15–19.

Jones, N, Ashton, K, and Chen, S-Y (2005) Rising to the Challenge of ASSET Languages, *Research Notes* 19, 2–4.

McNamara, T (1996) *Measuring Second Language Performance*, London: Pearson Education.

Milanovic, M and Saville, N (1996) Introduction, in Milanovic, M and Saville, N (Eds) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Studies in Language Testing, Vol. 3, Cambridge: UCLES/Cambridge University Press.

North, B (2004) *Europe's framework promotes language discussion, not directives*, *The Guardian*, retrieved from www.education.guardian.co.uk/tefl/story/0,5500,1191130,00.html

Weir, C (2004) *Limitations of the Council of Europe's Framework (CEF) in developing comparable examinations and tests*, paper presented at BAAL, King's College, London, 9–11 September 2004.

Can Do self-assessment: investigating cross-language comparability in reading

KAREN ASHTON, RESEARCH AND VALIDATION GROUP

Introduction

Asset Languages is a Department for Education and Skills (DfES) funded project to develop assessments for the DfES' Languages Ladder. At this stage of the Asset Languages project, assessments are available by skill (reading, listening, speaking and writing) across primary, secondary and adult contexts in eight languages with assessments currently being developed in a further 11 languages. Asset Languages assessments are constructed using the same format and task types across languages, which facilitates comparability to some degree, however issues still remain in ensuring that both the level of difficulty of assessments and the level of ability of the learners taking the assessments are comparable. This article discusses the development and piloting of a Can Do self-assessment tool to investigate the comparability of reading ability across learners of German, Japanese and Urdu. Before outlining the development of the self-assessment tool, sections on the need for cross-language comparability and current research on can do statements and self-assessments are provided.

The need for cross-language comparability

Cross-language comparability is an important research area for Asset. Having comparability across assessments in different languages provides a common discourse for candidates, language testers, educators, and employers. Scores from tests would thus be 'comparable across different languages and contexts' (Bachman and Clark, quoted in Bachman 1990: 5–6). As the *Common European Framework of Reference* (CEFR) states, this approach provides 'a sound basis for the mutual recognition of language qualifications' (Council of Europe 2004: 5). In the English educational system this kind of comparability across exams has, to date, not been addressed in a rigorous way. For example, Coleman (1996: 7) argues that labels such as 'first year level' or 'foreign language to degree level' are meaningless because of discrepancies in foreign language proficiency across English universities. Additionally, there has been criticism over the lack of comparability of grades in modern foreign languages for school examinations in 2005 (Halpin 2005).

Can Do statements

Can Do statements, e.g. *CAN understand short, simple texts if they use very common words or words that are easy to guess (e.g. postcards, timetables)* (CEFR portfolio statement, Ankara University 2000) are now commonly used for both teaching and assessment purposes. Little and Perclova (2001: 55) discuss how the use of Can Do statements 'encourages a generally positive attitude' in learners because as Cohen et al. (2000: 318) claim it focuses on the achievement of objectives and learning outcomes rather than on comparisons with other students. Can Do self-assessment tools give learners a sense of control and ownership over their learning (Little and Perclova 2001:53, Ushioda and Ridley 2002:42). The use of Can Do statements in teaching and assessment complements the increasing use of European Language Portfolios as a way of capturing students' work.

Despite the positive impact of Can Do statements on the motivation of learners, academics have criticised their use. In language testing, functional 'user-oriented' (Alderson 1990) statements are organised in a hierarchy which provides 'an operational definition of knowing a language' (Shohamy 1996:145). The benefit of this approach is that the functions are easy to comprehend, however, as Shohamy (1996:146) argues, the danger is that they give the 'illusion that they were based on something scientific, on theory'. The lack of theoretical basis of descriptors such as those in the CEFR has been critiqued, particularly their failure to offer 'a view of how language develops across these proficiency levels in terms of cognitive processing' (Weir 2004:5) and the fact that the scales present 'a taxonomy of behaviour rather than a development' in reading abilities (Alderson et al 2004:3). This criticism is also discussed at length by Mitchell (2003:5) who comments (with reference to the National Curriculum for Modern Foreign Languages) on the restrictive nature and ladder-like progression imposed by the Can Do statements.

Other criticisms of Can Do statements, particularly in relation to the CEFR have focused on the inconsistency and vagueness of the terms used, e.g. 'short', 'familiar' etc (Alderson et al. 2004:11–12). This is a fair criticism as many of the Can Do statements do exhibit vague language. However, in terms of using Can Do statements for learner self-assessment, it is perhaps of more relevance to find out how people, i.e. learners and teachers, interpret the statements rather than to critique them on this academic basis. Piloting and empirically evaluating Can Do statements, as was done for the CEFR, is one way of doing this. As Jones (2002:4) states 'frameworks by their nature aim to summarize and simplify, identifying those features which are common to all language users and all situations of use, in order to provide a point of reference for particular users and particular situations of use'. Arguably if learners have a shared understanding of the Can Do statements and are able to interpret them in a consistent manner, then the Can Do statements are fulfilling their function.

Self-assessments

Despite its growing popularity in the classroom, self-assessment has often been regarded as 'quite inappropriate for purposes of assessment' (Oscarson 1989:2). The main arguments against self-assessment have been that subjective estimates are inherently unreliable. Nevertheless, this article aims to demonstrate how learner self-assessments can usefully contribute to research in language testing. Taking research findings from previous studies into account can aid in the development of a more valid and reliable self-assessment tool and provide guidance as to the limitations in interpreting the findings.

Bachman and Palmer (1989:15) found self-ratings to be more reliable than they expected but as Ross (1998:6) discusses, in a review of studies, self-assessments correlate very differently across skills with reading correlating more strongly than listening, speaking and writing. Self-assessments of reading therefore tend to produce more valid results than self-assessments of the other three skills. Ross (1998:5) also notes that there tends to be considerable variation in learners' capacity to self-assess their ability in a second language. This finding was also reflected in the ALTE Can Do study where Jones (2002:13) found that correlations between self-ratings and exam grades were weakened when looking at individual responses. When correlating self-assessment data with exam grades, this highlights the need to look at mean ratings of self-assessments of candidates for each grade rather than analysing the data at the level of the individual.

Further studies in this area seem to confirm a pattern in the way that learners assess themselves. For example, it appears that overestimation is 'most evident for less experienced learners' and under-estimation for more experienced learners (Heilenman 1990:174, 190, Ross 1998:7, Jones 2002:15). This is likely to be because 'novice learners have little or no way of being aware of what they do not know or are unable to do. The more experience that learners have in a domain, however, the more likely they are to be aware of the limits of their skills and knowledge' (Heilenman 1990:190). In a study which contrasts with this pattern, Shameem (1998:104) found that despite a very strong correlation between learners' self-assessed ability and their ability as rated by an expert, learners were likely to marginally over self-assess at all levels. These results were based on a small sample of recent Fijian-Indian immigrants to Wellington, New Zealand and learners (predominantly in their early teenage years) were rating their L1 (Hindi). According to Heilenman (1990:189), learners evaluate their ability in terms of appropriateness and self-image before responding to self-report questions. This need to protect self-image, particularly considering learners were rating their L1, could account for the difference found in Shameem's study.

This need to protect self-image is also one difference in the way in which learners and teachers assess learners' ability. Additionally, learners tend to 'randomly walk through' their cognitive representations searching for a concrete example of what they can do whereas teachers' assessments are based on 'cumulative experience in observing student performance in the classroom context' (Ross 1998:16).

In terms of increasing the accuracy of self-assessments, Ross (1998:16) suggests using assessments based on functional skills that learners can relate to rather than abstract skills that learners are likely to have had less direct experience of. In addition to this, Jones (2002:15) discusses how longer statements or paragraphs produce a more discriminating tool than short statements as short statements 'do not "epitomize" levels' in the same way'.

The next section discusses the development of a Can Do self-assessment survey for reading.

Piloting a self-assessment tool

A Can Do self-assessment survey for reading, intended for 12–14 year-old secondary school learners of Urdu, German or Japanese in England, was developed as part of a pilot study. Given the age of the learners, the survey needed to be easy to use and interpret. It was also important that the survey discriminated between learners of different levels of ability and could be used for learners of the various languages under investigation. In addition to the learner self-assessments, teachers were asked to give ratings for the learners and provide additional information such as National Curriculum levels.

Rather than creating new Can Do statements, North's (2000:182) procedure of using existing scales to develop the Can Do survey was followed. A range of statements were reviewed that:

- had been calibrated, e.g. CEFR (Council of Europe 2001) and ALTE Can Do Project (ALTE 2002)
- have been used for this age group, e.g. Bergen Can Do Project (Hasselgreen 2003) and CEFR portfolios adapted by different countries for 11–16 year olds
- English learners are familiar with, e.g. Languages Ladder Can Do statements (Asset Languages 2005) and the National Curriculum for England: Modern Foreign Languages statements (Department for Education and Employment and Qualifications and Curriculum Authority 1999).

Sorting exercises were performed by the researcher where statements were grouped according to salient features. Where statements were the same or very similar, the statement with the most concrete example was chosen. Where possible, preference was given to statements from the CEFR or the ALTE Can Do Project as these statements have been calibrated empirically. Although longer statements tend to discriminate more, given the age range of learners and the need for the survey to be completed relatively quickly (typically during part of a lesson), short statements were piloted. This compromise was felt necessary given these constraints and the importance of learners not feeling daunted by the appearance of the survey.

The survey, containing a total of 43 statements across three levels (approximately CEFR levels A1–B1), was piloted on 124 secondary learners of German, Japanese and Urdu. Although learners were expected to be at A1 or A2 level, it was important to have statements at B1 in case some learners were at a higher level than expected and to discriminate at the higher end of the scale. Following the format that Jones (2000) used in the ALTE Can Do Project, learners were asked to select 'yes', i.e., 'I can' or 'no', i.e., 'I can't' for each Can Do statement. For ease of completion the survey was divided into five sections. Two versions of the survey were piloted. Version A had the Can Do statements in order of predicted difficulty (from easiest to most difficult) while version B had the statements in random order within each section. The purpose of this was to test whether having the statements in order of perceived difficulty provided extra scaffolding for interpretation of the statements. Version A takes into account the findings of Bachman and Palmer (1989) that learners are more aware of what their difficulties are and are better able to assess what they cannot do rather than what they can do. If the order provides scaffolding for learners, they should be able to more clearly locate the cut-off point between what they can do and what they cannot do.

Results

There was a high correlation between version A and version B of the survey showing that learners were rating the statements in the survey in very similar ways across both versions. One main difference was that Version A discriminated better than Version B which is intuitive given that the statements were provided in predicted order of difficulty and learners could more easily 'draw the line' between statements that they could or could not do. Version A, with Cronbach's alpha of 0.88 had slightly higher reliability for the Can Do statements than Version B with Cronbach's alpha of 0.83. Given these findings, it was decided to use version A of the survey for future data collection sessions.

Initial analysis was performed separately for the three languages and misfitting persons and statements were removed from the data set. From 43 statements, seven were removed during this process due to poor fit values. In this case, the poor fit values signified that they were redundant statements, poorly worded, poorly discriminating, or were measuring the same ability as other statements in the survey (McNamara 1996:176). A subsequent analysis was performed combining the data sets. High reliability with a Cronbach's alpha of 0.94 for the Can Do statements was achieved from this analysis. Table 1 shows the correlations between the difficulty values of the Can Do statements for the three languages.

Table 1 : Correlations between difficulty of Can Do statements

German and Japanese	German and Urdu	Japanese and Urdu
0.81	0.81	0.66

This table shows that the German data correlated well with both the Japanese and Urdu data with a weaker correlation for the Urdu and Japanese data. All of these correlations are statistically significant at the 0.05 level of significance, i.e. there is 95% confidence that these findings are not due to chance. The weaker correlation showing greater discrepancy between the way Japanese and Urdu learners interpreted the statements will be investigated further with a larger number of candidates.

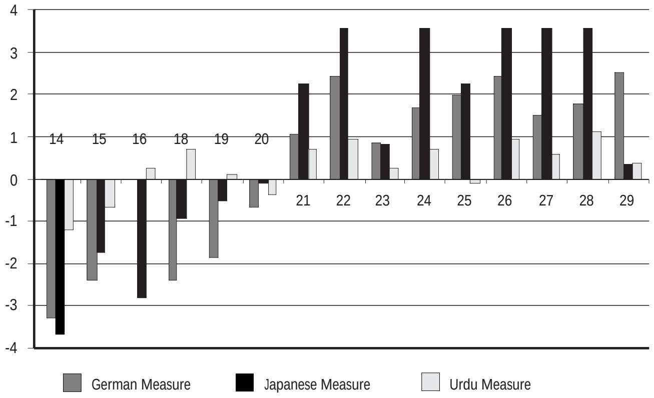

Figure 1 : Difficulty values of statements across languages

Figure 1 shows the difficulty values of the middle 15 statements across the three languages. This figure shows that the majority of the statements have calibrated according to expectations and have worked similarly across languages. Within each grouping, the third of four being shown in Figure 1, it can be seen that the starting statements have very low difficulty values (i.e. are easy statements e.g. statement 14) and that these get progressively higher (i.e. are more difficult statements), towards the end of each scale (statement 29). There were also some differences in the way that statements calibrated and fitted the model for each language. For example:

> *CAN understand the main points of the news in newspapers and magazines (statement 25)*

This statement came out as much more difficult than anticipated for Japanese learners and easier than expected for Urdu learners. According to the Japanese teacher, fulfilling this function in Japanese would require the knowledge of over 1000 kanji characters[1] meaning that the level of this statement for Japanese learners is likely to be above the level of its given CEFR level of B1.

Another statement which calibrated differently to expectations is:

> *CAN read books if I am very motivated and they use familiar language (statement 29)*

This statement calibrated as easier than expected for both Urdu and Japanese learners. Both of these groups of learners read simple books in the classroom to practise reading the script. This Can Do statement is from the CEFR where it is likely that the intended interpretation was that of a novel.

The difference in the way that these statements have calibrated highlights the issue that the CEFR is designed for latin script languages and does not address issues of script acquisition for languages such as Japanese or Urdu. This issue and further differences in the difficulty of functions across languages will be explored in more depth with a larger number of learners in future data collection sessions.

In order to determine how well the survey discriminated across levels, it was necessary to estimate the degree of mastery needed to say that a learner was at a given level. Following Hasselgreen (2003:28) and Jones (2000:13), 80% was used to determine whether learners are 'operating at the level'. Using this value, the analysis shows that 85 of the 124 learners would have 80% chance of success at A1, 26 at A2 and 13 at B1. The average difficulty values for statements at each level are shown in Table 2. This shows that learners in all three languages rated statements in a way that showed a consistent progression from statements at A1 to A2 to B1.

The survey was also able to discriminate levels of different language learners and showed that the average ability of this group of Urdu learners was marginally higher than the average ability of German learners. Both of these groups had average abilities at least a level above the Japanese learners.

Table 2 : Average ability values for levels

Level	German	Japanese	Urdu
A1	-2.15	-2.00	-0.83
A2	0.45	0.73	0.32
B1	1.86	1.35	0.71

In terms of the accuracy of self-ratings, the teacher ratings were consistently more modest than learner ratings for both the German and the Japanese learners showing that it is likely that both of these groups have over-rated their ability as is common of lower proficiency learners. The Urdu teacher ratings did not correspond well to the learner ratings and it is therefore difficult to determine whether the Urdu learners have over-rated their ability. If it could be assumed that lower proficiency learners of all three languages over self-assess themselves to the same extent, then this issue would be cancelled out and comparisons easy to make across languages. This assumption and possible reasons for the lack of correspondence of the Urdu teacher and learner ratings need to be explored further.

Conclusion

This article has shown that self-assessments are a useful tool in investigating the ability levels of learners of different languages. Further self-assessment data is currently being collected using the tool developed and discussed here. The same candidates completing the self-assessments are also sitting Asset Languages assessments in these languages. Candidate grades will act both as a validity check for the self-assessment and to determine the extent of over- or under-rating. However, as Asset Languages is in the early stages of scale construction and test development, the levels are being closely monitored and are subject to further fine-tuning. As Upshur (quoted in Heilenman 1990:174) states 'learners have access to the entire gamut of their success and failures in the use of the second language whereas any test of actual language use, of practical necessity, can sample only a small proportion of that ability'. The self-assessments can therefore provide additional information on the ability of learners which can feed into future test construction.

[1] Japanese has three character systems – kanji, hiragana and katakana. Kanji characters are derived from Chinese characters, although now many of these characters have been modified in Chinese whereas traditional forms have been kept in Japanese. In 1981, the Japanese Government issued a list of 1945 characters recommended for daily use.

References and furher reading

Alderson, C (1990) Bands and Scores, in Alderson, C and North, B (Eds) *Language Testing in the 1990s*, London, British Council/Macmillan, Developments in ELT.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S, and Tardieu, C (2004) *The Development of Specifications for Item Development and Classification within the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Reading and Listening. The Final Report of the Dutch CEFR Construct Project*, Project Report, Lancaster University, Lancaster, UK, retrieved from: www.eprints.lancs.ac.uk/44/

ALTE (2002) *ALTE can-do project*, retrieved from www.alte.org/can-do/index.cfm

Ankara University (2000) European Language Portfolio for 11–15 year old learners, Tömer, Ankara University.

ASSET Languages (2005) *Can-Do statements*, retrieved from www.assetlanguages.org.uk

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

Bachman, L F and Palmer, A S (1989) The construct validation of self-ratings of communicative language ability, *Language Testing* 6 (1), 14–29.

Cohen, L, Manion, L and Morrison, K (2000) *Research Methods in Education*, London: Routledge-Falmer.

Coleman, J (1996) *Studying Languages: A survey of British and European students*, London: CILT.

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

—(2004) *Plurilingual Education in Europe: Draft 1*, paper presented at Global Approaches to Plurilingual Education, Strasbourg, 28–29 June 2004.

Department for Education and Employment and Qualifications and Curriculum Authority (1999) *The National Curriculum for England: Modern Foreign Languages*, London: Department for Education and Employment, Qualifications and Curriculum Authority.

Halpin, T (2005) *Exam watchdog reviews GCSE glitches in languages grading*, retrieved from http://www.timesonline.co.uk/article/0,,2-1742559,00.html

Hasselgreen, A (2003) *Bergen Can Do Project*, Strasbourg: Council of Europe.

Heilenman, K (1990) Self-assessment of second language ability: the role of response effect, *Language Testing* 7 (2), 174–201.

Jones, N (2000) Background to the validation of the ALTE 'Can-do' project and the revised Common European Framework, *Research Notes* 2, 11–13.

—(2002) Relating the ALTE Framework to the Common European Framework of Reference, in Council of Europe (Eds), *Case Studies on the use of the Common European Framework of Reference*, Cambridge: Cambridge University Press, 167–183.

Little, D and Perclova, R (2001) *European Language Portfolio Guide for Teachers and Teacher Trainers*, Strasbourg: Council of Europe.

McNamara, T F (1996) *Measuring Second Language Performance*, Harlow: Longman.

Mitchell, R (2003) Rethinking the concept of progression in the national curriculum for modern foreign languages: A research perspective, *Language Learning* 27, 15–23.

North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang.

Oscarson, M (1989) Self-assessment of language proficiency: rationale and applications, *Language Testing* 6 (1), 1–13.

Ross, S (1998) Self-assessment in second language testing: a meta-analysis and analysis of experiential factors, *Language Testing* 15 (1), 1–20.

Shameem, N (1998) Validating self-reported language proficiency by testing performance in an immigrant community: the Wellington Indo-Fijians, *Language Testing* 15 (1), 86–108.

Shohamy, E (1996) Competence and performance in language testing, in Brown, G, Malmkjaer, K and Williams, J (Eds) *Performance and Competence in Second Language Acquisition*, Cambridge: Cambridge University Press, 136–151.

Ushioda, E and Ridley, J (2002) *Working with the European Language Portfolio in Irish Post-Primary Schools: Report on an Evaluation Project*, Trinity College Dublin, CLCS Occasional Paper No. 61.

Weir, C (2004) *Limitations of the Council of Europe's Framework (CEF) in developing comparable examinations and tests*, paper presented at BAAL, King's College, London, 9–11 September 2004.

Assessment processes in Speaking tests: a pilot verbal protocol study

CHRIS HUBBARD, PERFORMANCE TESTING UNIT
SUSAN GILBERT AND JOHN PIDCOCK, SENIOR TEAM LEADERS

Introduction

There has been growing interest, over recent years, in the processes involved in the assessment of candidate performance, but few studies have directly investigated how raters make their assessments in 'real time'. Work has either focused on other issues related to Speaking test assessment such as the question of inter-rater reliability (Adams 1978 cited in Fulcher 2003, 140), comparison of the ratings of individual examiners against each other, thereby identifying rater severity and consistency (McNamara 1996), or has used retrospective data captured after the Speaking test event, not as the assessment is taking place (Orr 2002). Studies using 'real time' Verbal Protocol Analysis (VPA) techniques have focused on aspects of the assessment of writing, such as the decision-making processes employed by raters (Milanovic, Saville and Shen 1996), and how raters interpret and apply rating scales (Falvey and Shaw 2006).

The work described here sets out to answer questions related to the decision-making processes of CAE Speaking test examiners as they use the assessment scales in 'real time'. The key objective of this pilot stage is to investigate the following research questions:

- Can useful data be captured from Speaking test examiners in 'real time'?
- What do the data indicate about how examiners use rating scales?
- What aspects of performance are focused on during individual parts of a test, and are these similar across tests?

The aim is not only to gain a better understanding of how examiners approach their work, but also to use that understanding to inform examiner training and development programmes, and to provide feedback to examiners on how they use the scales.

Methodology

Qualitative research methods have an important and growing role in research into the assessment of performance skills. Taylor (2005:3) outlines developments in the use of qualitative research methods at Cambridge ESOL, including Verbal Protocol Analysis, to provide 'rich insights, not only into test-taker processing, but also into the attitudes and behaviour of … examiners.'

A VPA methodology was selected for this study. In VPA approaches participants verbalise their thoughts, either during the assessment process or immediately afterwards. In order to capture the process of assessment 'as it happens' a 'Think aloud' procedure without mediation was chosen (see Green 1998 for a discussion of the options in VPA research). In order not to affect the performance of the candidates, the examiners rated video recorded performances. 'Thinking aloud' required them to verbalise all heeded information as they watched the video recordings and carried out the process of assessment in the normal way. Each examiner worked individually; they did not discuss their thoughts with and were not prompted by a mediator.

The validity of verbal reports is crucial to the capture of useful data and to this end a focus was placed on ensuring reports were as complete as possible and that they were captured with minimal disruption to the assessment process. Trialling of the methodology helped to refine the approach, and also served to clarify with the participants that the study would not make strict demands in terms of what the examiners were expected to verbalise. They were asked only to 'think aloud' as they reacted to and assessed candidate performances.

The study

The study was set up as part of a routine exercise carried out by Cambridge ESOL. In this exercise groups of senior and experienced examiners watch and rate around 10 Speaking tests from a particular examination on video in order for representative and dependable scores to be produced. These videos and accompanying scores are later used in the training, standardisation and co-ordination of Oral Examiners. The three examiners involved in this study were members of the team scoring CAE performances and were involved in rating other Cambridge ESOL tests. This was important as it allowed us to verify that the ratings these examiners made of the candidates involved in the study were consistent with those they made for the remaining seven CAE tests and with their performance on other tests: that engaging in the protocol study did not make them harsher or more lenient as raters than they would otherwise be. Using a Multifaceted Rasch approach it was possible to compare estimates of intra- and inter-rater reliability for these examiners. None was found to be misfitting (inconsistent with the ratings given by the group of examiners as a whole) either on the CAE tests or on the other tests in which they were involved.

The three participating examiners were asked to verbalise their thoughts for three of the ten CAE tests; tests 3, 6 and 9 in the sequence. For each test a variation on the basic data capture model was introduced. We set out to evaluate these three variants on the think-aloud model as a basis for recommending a single 'best model' for use in future studies. The variants were:

1. **Test 3**, Think aloud, no separate notes to be made, the test to be played and the think-aloud carried out without pausing the video.
2. **Test 6**, Think aloud, separate written notes could be made if examiners would usually do so, the test to be played and assessed without pausing the video.
3. **Test 9**, Think aloud, the test could be paused at the end of each part of the test if the examiner wished to add relevant comments about the performance, separate written notes could be made.

While assessing, the examiners recorded on cassette or mini disc any thoughts they verbalised during each test. After completing all the tests the recordings were sent to Cambridge and transcribed. For each test, the three examiners' comments were transcribed in parallel columns. Interlocutor and candidate speech columns were also included to create an indicative framework showing when during each test the comments had been made. A system of coding was then agreed by the researchers and the protocols coded.

Coding the data

As described by Green (1998:12) the 'validity of codings is related to the reliability of codings'. To support meaningful conclusions, the process of coding is key. For this investigation the following categories were identified:

- Grammar and Vocabulary (GV)
- Discourse Management (DM)
- Pronunciation (P)
- Interactive Communication (IC)
- Assessment Comment/Decision (ASS C/D)
- Other

The first four categories directly reflect the CAE analytical assessment criteria. The Assessment Comment/Decision category was used where examiners mentioned a likely score on the five-band CAE scale, such as 'She's not up to 5 but she's better than 3' or commented on how the performance may affect the assessment decision, e.g. 'It's better than adequate but it's not top range'. The Other category was used to code any utterances not directly covered by the assessment categories above, such as details about the recording or test process; 'OK, they're starting part 2 now', affective comments such as; 'it's a shame V didn't respond there' and any other speech not covered by the five main categories. Although the discussion below does not expand on the nature of comments in the Other category, they made up approximately 16% of the overall comments and it is planned to look at these in more detail at a future date.

All the transcripts were independently coded by the three researchers and then discussed by the group to clarify any disagreements that had arisen. The independent coding process returned a high degree of agreement between coders when allocating comments to the six categories listed above. Issues that needed clarification centred around extended comments and the number of coding categories to be assigned. It had initially been an intention to further categorise comments for each assessment criterion according to the relevant 'sub-criteria', e.g. for GV the sub criteria of Range, Accuracy and Appropriacy. However, there were clear differences in interpretation between the three researchers in coding some sub-criteria. For example, the examiner comment, '"*cut it all away, the fores*t" *that's candidate A*', was unanimously assigned to the GV category, but the researchers disagreed as to whether it should be further categorised as *range*, *accuracy* or *appropriacy.* When coding 'hesitation' there was also some discussion about whether an examiner was referring to a candidate's ability to manage the discourse or their ability to interact with others. This aspect is currently covered under IC and was therefore coded as such following the discussion. There were similar issues with the sub-criteria for all four analytical scales. This led the researchers to settle on the six relatively broad coding categories listed above. The discussion also produced suggestions for reviewing and clarifying descriptions and instructions for the application of the sub-criteria, and Cambridge ESOL is currently investigating these issues with a wider sample of Oral Examiners.

Findings

The findings from this study are presented here in relation to the three questions outlined at the start of this article.

Can useful data be captured from Speaking test examiners in 'real time'?

All three VPA models described above generated data that:

- were captured in 'real time' without any apparent detrimental affect to the assessment process
- could be coded consistently using a standard coding model
- appear meaningful.

Possible limitations of the VPA approach need to be borne in mind when answering this question more fully, such as the general assumption that verbal reports give an insight into cognitive processes, and, particularly relevant to this study, the extent to which the 'think aloud' procedure may affect the actual process it is reporting on, i.e. assessment. However, there can be few research tools without limitations and care has been taken in this study to minimise restrictions of the methodology by using video performances, asking examiners to only verbalise their thoughts as they assess and employing the validation checks on the examiner assessments as part of a larger routine exercise.

Overall, VPA appears to be a credible way of capturing useful data in 'real time'. In this study, model 3 above, where examiners could stop the video of the test to record extra assessment-related comments, was least preferred by the group as it differed too much from the routine process of assessment. It also led to examiners commenting at more length during breaks in the test, and less during actual assessment, which detracted from the aim of 'online' data capture. The recommended approach for future work would be to give examiners an option of model 1 or 2.

What do the data indicate about how examiners use rating scales?

Findings in relation to this question have been divided into two specific areas that arose from the analysis:

How much do examiners focus on the different analytical scales?

Over the three tests the total number of comments relating to each of the four assessment scales and to assessment comments or decisions are shown in Figure 1.

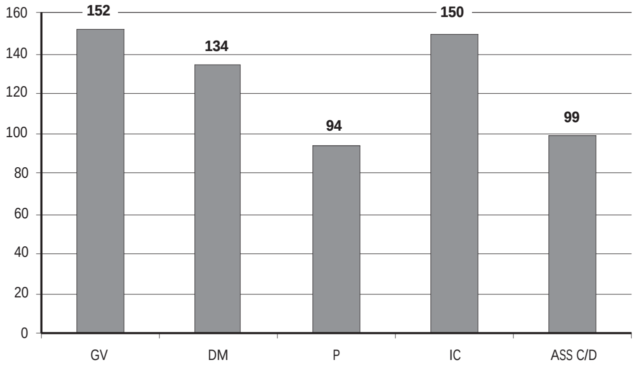

Figure 1 : Number of comments relating to each assessment scale

There were considerably fewer comments on Pronunciation (P) than the other three assessment scales. This could be because examiners seem to make decisions about pronunciation relatively early in the test (see below) and may feel that it remains fairly constant throughout. GV and IC were commented on almost equally while DM featured slightly less overall. Although this may indicate that these examiners give greater attention to GV and IC, it could equally be the case that the more concrete or objective nature of aspects of the GV and IC scales, like 'grammatical accuracy' or 'initiating and responding', make them intrinsically easier for raters to identify and comment on spontaneously. This may also be a shortcoming of the protocol approach in that it only captures the most 'conscious' or easily articulated aspects of the rating process. It is conceivable that there are other less accessible processes that are not being captured.

How do examiners appear to be using and interpreting the sub-criteria within each criterion?

The data from the first test was coded and analysed in more detail, concentrating on use of the sub-criteria. As mentioned above, coding the comments to the various sub-criteria proved problematic, indicating that they could, in practice, be open to various interpretations; despite this limitation, it was possible to produce some tentative findings.

Taking all three examiners together for test 1, all of the sub criteria were referred to at some point during the test, but there were very strong peaks and troughs, as shown in Figure 2.

Figure 2 : Number of comments per sub-criteria for all three raters

Sub-criteria	Comments
General	47
GV general	6
GV range	31
GV accuracy	18
GV appropriacy	6
GV assessment scales	2
DM general	7
DM relevance	3
DM coherence	20
DM extent	21
DM assessment scales	7
P general	10
P stress and rhythm	5
P intonation	1
P individual sounds	8
P assessment scales	5
IC general	12
IC initiating and responding	33
IC hesitation	14
IC turn-taking	8
IC assessment scales	8
Assessment Scales	3
Assessment comment/decision	49
Level	5

From this limited amount of data, the initial findings suggest the following points about each criterion:

- **Grammar and Vocabulary** (GV) – 'range' was noticed more than 'accuracy'. 'Appropriacy' was referred to significantly less, perhaps indicating overlap between appropriacy and the other GV concepts in examiners' minds.
- **Discourse Management** (DM) – 'coherence' and 'extent' were noticed much more often than 'relevance', which may be more difficult for examiners to identify precisely in the speaking test scenario.
- **Pronunciation** (P) – 'individual sounds' were commented on more than 'stress and rhythm', and 'intonation' was hardly ever referred to. In fact comments coded as 'P general' made up the biggest group for pronunciation. Remarks such as '*Both (candidates) quite clear*' and '*very good pronunciation there*' tended to suggest that examiners approach this criterion from a more holistic reaction to 'comprehensibility' and 'strain on the listener', which are both elements of the detailed descriptors that make up the assessment scales.
- **Interactive Communication** (IC) – 'initiating and responding' was commented on by far the most from the IC sub-criteria. A clear boundary between 'turn-taking' and 'initiating and responding' as separate criteria may not always be evident to examiners.

What aspects of performance are focused on during individual parts of a test?

Figure 3 shows the percentage of comments made for each of the four assessment scales by test part, for all three tests.

Figure 3 shows that features of performance from all four scales were referred to in all four parts of the test, but not equally. The assessment comment/decision category is also included and again has an individual profile. But were these patterns similar across tests?

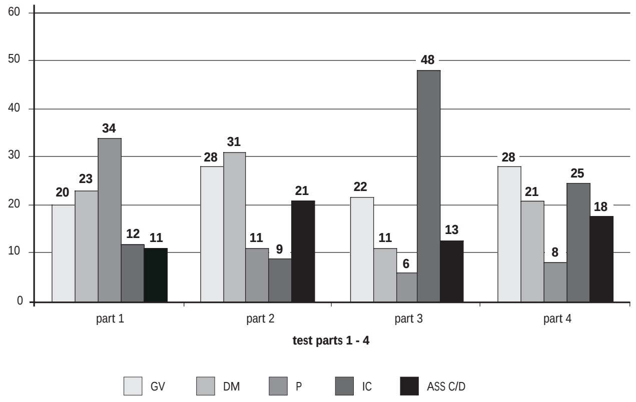

Figure 3 : Percentage of comments per scale by test part

When analysed separately, this pattern of distribution of comments was closely mirrored across all three tests. This is shown in Figure 4 where each criterion is ranked from 1–5 for each test according to the percentage of comments received in each part of the test. The criterion ranked 1 received most comments. The ranking index (RI) is a simple total of the rankings on each of the three tests with a possible range from 3 (i.e. for an item ranked first on each of the three tests) to 15.

Although there are differences between individual tests it is clear that some general trends can be identified:

- Grammar and Vocabulary was commented on at a fairly constant level throughout the test with ranking indexes between 5–7 for each part.
- Discourse Management received a good proportion of the comments in parts one and two (RIs 4,5), but significantly less in part three (RI 10).

Figure 4 : Ranked distribution of comments across parts of three tests

	Criterion	Ranking			Ranking Index
		Test 1	Test 2	Test 3	
Part 1	GV	3	2	2	**7**
	DM	1	2	2	**5**
	P	2	1	1	**4**
	IC	4	5	4	**13**
	ASS C/D	5	4	5	**14**
Part 2	GV	1	2	2	**5**
	DM	2	1	1	**4**
	P	5	5	4	**14**
	IC	4	4	5	**13**
	ASS C/D	3	3	3	**9**
Part 3	GV	2	2	3	**7**
	DM	3	3	4	**10**
	P	5	5	5	**15**
	IC	1	1	1	**3**
	ASS C/D	4	3	2	**9**
Part 4	GV	3	2	1	**6**
	DM	3	4	2	**9**
	P	5	4	5	**14**
	IC	2	1	3	**6**
	ASS C/D	1	3	3	**7**

- Pronunciation was key in part one of the test, with an RI of 4, and much less so thereon (RIs 14,15) – this could suggest that examiners make a general decision on pronunciation quite early on, perhaps feeling it is likely to remain fairly constant throughout the test.
- Interactive Communication was clearly the focus in part three with an RI of 3.
- In part four, the comments are fairly equally distributed across the GV, DM and IC scales, suggesting perhaps that this stage is used by examiners to check or finalise their grades. This may be supported by the fact that there is far greater variety in ranking for each criterion across the three tests with only P never being ranked 1 or 2.
- Assessment Comment/Decision is ranked fairly constantly from part 2 onwards (RIs 9,7) suggesting a constant review of examiner decisions throughout the test. It should also be noted that there is more variation in Ass C/D rankings in parts 3 and 4 of the test perhaps suggesting that variations in individual performances could affect when examiner decisions are formed. However, this is an area that needs more focused investigation.

Obviously, the nature of the tasks in each part of the test will contribute to what examiners focus on. For example, aspects of IC are much more a feature of the collaborative tasks used in CAE parts 3 and 4, than those used in parts 1 and 2. But what does seem significant is that the pattern of focus was reproduced by all three examiners over the three tests.

Conclusion

This study set out to gather ‘real time’ data from, and insights into, the processes involved in the assessment of Speaking test performances, and as an initial stage in that process has, on the whole, been successful. However, it does need to be seen as a preliminary step. Both the evaluation of the research design and the applicability of the findings would benefit from the replication of the current research in larger scale studies. Fresh questions and lines of enquiry have inevitably arisen from this experience that it would be interesting to pursue in further research. For example, examiners involved in this study were found to be focusing on different aspects of performance at different stages during the test. This is not encouraged in examiner training. In fact examiners are instructed to consider all elements throughout the whole test. It is therefore reasonable to assume that this approach to rating must have developed independently for each of the examiners involved. Questions arising from this observation are: Is it in fact possible to train examiners to follow current guidelines and apply four scales continuously throughout a test? Is the differential attention to criteria due to features of the assessment scales and aspects of the language elicited by particular tasks, or is it more fundamental to the nature of the assessment process?

Although we acknowledge their limitations, we believe that the data collected in this pilot study may nonetheless provide further insights into how the CAE rating scales are being used. Assessment comments and decisions appear to be prominent in examiners’ minds fairly consistently from Part 2 of the test, so analysing these for further insights into the process of arriving at scores is an intention. We also intend to report on the content of comments coded to the ‘Other’ category.

Overall, this has been a useful and productive study in terms of selecting and trialling a suitable data capture model and presenting initial findings. It has also served to focus thoughts on salient features of the assessment process for consideration in further Oral Examiner training and development programmes. It will have applications beyond the assessment of speaking and could, for example, be used to inform the assessment of other skills such as Writing.

References and further reading

Falvey, P and Shaw, S (2006) IELTS Writing: revising assessment criteria and scales (Phase 5), *Research Notes* 23, 8–13.

Fulcher, G (2003) *Testing Second Language Speaking*, Harlow, Pearson Education Ltd.

Gass, S, M and Mackey, A (2001) *Stimulated Recall Methodology*, Mahwah, New Jersey: Lawrence Erlbaum Associates.

Green, A (1998) *Verbal protocol analysis in language testing research: a handbook*, Studies in Language Testing, Vol. 5, Cambridge: UCLES/Cambridge University Press.

Luoma, S (2004) *Assessing Speaking*, Cambridge, New York: Cambridge University Press.

McNamara, T (1996) *Measuring Second Language Performance*, London, New York: Longman.

Milanovic, M, Saville, N and Shen, S (1996) A study of the decision-making behaviour of composition markers, in Milanovic, M and Saville, N (Eds) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium*, Cambridge and Arnhem, Studies in Language Testing, Vol. 3, Cambridge: UCLES/Cambridge University Press.

Orr, M (2002) The FCE Speaking test: using rater reports to help interpret test scores, *System* 30, 143–154.

Taylor, L (2005) Using qualitative research methods in test development and validation, *Research Notes* 21, 2–4.

IELTS Writing: revising assessment criteria and scales (Conclusion)

STUART SHAW, RESEARCH AND VALIDATION GROUP

Introduction

Over the last four years a series of articles describing the five phases of the IELTS Writing Assessment Revision Project have been published in *Research Notes*: *Initial Planning and Consultation* (Shaw 2002a); *Development* (Shaw 2002b; Shaw 2004c); *Validation* (Shaw 2004d); *Implementation* (Bridges and Shaw 2004a); and *Operation* (Shaw and Falvey 2006).

The articles have endeavoured to chronicle and discuss the processes and outcomes of the project at each stage of its development prior to the operational implementation of the revised rating scale in January 2005. These articles have revealed what considerations 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.

The purpose of this final article in the series is to focus on the qualitative dimension of the final operational phase of the project and in particular to describe the findings from a global survey of key personnel at a number of IELTS Test Centres. Qualitative studies, which have featured prominently throughout the revision project, are often used to supplement the findings of large scale quantitative studies and enable researchers to delve deeper into subjects' minds (in this case of the revision project, the minds of examiners/administrators and candidates) and to elicit attitudes and opinions by allowing the subjects to explore topics that concern them through written or verbal means.

Surveys undertaken throughout the revision of the IELTS writing rating scale

Findings from a number of surveys conducted during the project are given here in summary form in order to demonstrate the iterative approach to test revision undertaken, that is, the extensive consultation with stakeholders; and the iterative procedures that occurred extensively, to amend and change assessment criteria, descriptive bands and the rubric.

Preliminary results of the feedback provided in an examiner questionnaire distributed globally during Phase 1 of the project provided many insights into what was then current assessment practice and highlighted the need for the revision of bands and descriptors, regular standardisation and training of examiners and the need for the production of effective training materials. The full findings are reported in Shaw (2002c).

In order to complement the quantitative Phase 3 Validation trial (reported in Shaw 2003) a qualitative trial was undertaken with four UK and eight Australian senior IELTS examiners. The areas previously identified during the Development Phase – assessment approach; assessment criteria; rating scale descriptors; examiner training – informed construction of the trial questionnaire. The trial demonstrated that examiners found the revised rating scale user-friendly and were, in general, favourably disposed towards the new rating approach and conscious of its potential. One principal concern relating to examiner re-training was whether examiners would be given sufficient time and opportunity to practise on a range of quality exemplar scripts. Some difficulties were envisaged due to standardising examiners with the new rating scale – practice time, text processing, and scale application. Full findings from the survey are in given in Shaw (2003a).

A short examiner training questionnaire was also distributed to senior examiners during the training phase (Phase 4) of the project in June 2004. The IELTS Chief Examiner in Australia and 13 senior examiners and trainers took part in the first training session held in Sydney. One chief examiner and five senior examiners/trainers took part in a subsequent UK training event. Full findings from feedback given at the first training sessions are reported in Shaw (2004a).

In 2004 training was cascaded from senior trainers to trainer trainers who in turn cascaded training to trainers then examiners. Feedback from several sessions involving trainer trainers together with some findings from sessions comprising examiner trainers is reported in Shaw (2004a). Comparisons were reported across the range of training personnel (including senior examiners from previous training sessions) where it was believed they would be insightful. Comments revealed a very favourable response to the revised rating scale and the Examiner Training Materials. The new band descriptors were seen as being clearer and much more user-friendly. It was believed that the new rating scale alongside the improved training, certification, monitoring and standardisation procedures under the auspices of the *Professional Support Network* (PSN) would ensure greater reliability of rating (see Mitchell Crow & Hubbard 2006).

There was another survey instrument used in 2005, the Phase 5 global survey. A preliminary trial of a questionnaire to be used for the global survey was undertaken in April 2005 with a group of IELTS writing raters from Anglia Ruskin University, Cambridge, UK (ARU) in order to ascertain how well the revised assessment criteria and band level descriptors were functioning. Findings are reported in full in Shaw (2005b) and summarised in Falvey and Shaw (2006).

Trial methodology

Two questionnaires, revised in the light of the ARU pilot trial, were constructed: an examiner questionnaire and a centre administrator's feedback form (in order to elicit views on the new scales from an assessment, administrative and practical point of view). The questionnaire was circulated to the top thirty IELTS test centres based on candidate entries. Centres were located in several continents including Europe, Australasia and Asia. An additional three UK centres were also targeted. Twenty examiner questionnaires were sent to each Test Centre (a total of 660 questionnaires). Each Test Centre Administrator also received a feedback form.

271 examiners returned completed questionnaires. It should be noted that the 41% return rate does not reflect the real percentage of examiner responses because a standard number of copies (20) were sent to each centre even though some centres have only five or six examiners. This situation roughly reflects the 211 responses to the questionnaires that were received in 2001 during Phase 1. Centre return rates varied between 100% and 0%. Only four centres failed to return in time for data analysis. Of the 33 *Administrator Feedback* forms which were despatched to test centres, 17 were returned. This represents a 52% return rate.

The examiner cohort consisted of a range of experienced EFL/EAP teachers and examiners. The background information of participating examiners is shown in Table 1.

The average number of years as an EFL/EAP teacher of the IELTS examiners in the trial is 14.91 years of which an average 5.44 years has been spent examining. Interestingly, a small proportion of participants appeared to have examining experience but no apparent history of teaching.

Findings

The feedback from examiners and administrators can be considered separately.

Feedback from Examiners

It was clear that an overwhelming majority of examiners appreciated the revised rating scale believing it to be a considerable improvement overall on the former one. General feedback from examiners was very positive. The need for a substantial revision of the IELTS rating scale had been broadly welcomed, the revision being regarded as both timely and considerable by examiners and administrators alike. IELTS examiners acknowledged that the revision project had been well-researched and empirically-grounded. Moreover, use of the new writing criteria engendered an excellent washback effect on assessment within certain language teaching institutions.

Table 1 : Examiner background information

Experience as an examiner

- Broad experience of Main Suite (Upper and Lower) – both oral and written examinations including pre- and post-revisions;
- Extensive involvement with other Cambridge ESOL, non-IELTS examinations such as BEC, BULATS, CELS, YLE, Skills for Life;
- Active participation in Cambridge CELTA and DELTA awards;
- Extensive range of IELTS examining including preparation course and tertiary-level sessional teaching;
- Widespread experience with other examinations, e.g. IGCSE; Language Benchmark Tests, TOEFL, RSA, CCSE.
- Pre-University General English;
- Undergraduate and postgraduate English examining for BA, BSc and BA (Lit), MA English Language exams, MPhil and PhD English Language.

Years as an IELTS examiner

Av = 5 Max = 35 Min = 8 months

Years as an EFL/EAP teacher

Av = 15 Max = 38 Min = 0

The new scale, it is believed, now offers examiners a good indication regarding where to place candidates on the IELTS proficiency continuum. The scale is perceived to be more helpful than the previous one, offering better guidance to examiners, and is considered to be fairer to the candidate. The changes made are believed to facilitate more efficient and effective marking engendering greater confidence amongst examiners and the new assessment scales appear to have reduced marking subjectivity.

Examiners were appreciative of the increased explanatory text accompanying the new descriptors as the revised text has allowed for 'greater delicacy of assessment'. Many ambiguities in the previous set of descriptors (such as the extremely short but word-perfect answer) have been cleared up and are now well-defined. The new band descriptors are considered to be more rigorous compared to their former counterparts. The descriptors are also clearer, more comprehensive, easier to follow and achieve a greater precision than before. Regarded as methodical, especially in relation to assessment approach and the clarity of word count, the revised descriptors are thought to be more accurate and easier to use as a discriminating instrument.

Examiners also felt that the new criteria helped a great deal with the problem of marking memorised or potentially memorised scripts though this still remains an area of some concern. The revised scale appears to deal very effectively with the problem of candidates supplying 'off topic' responses. The introduction of ceilings and penalties and the inclusion of descriptors legislating for the use of formulaic language, appropriate paragraphing and punctuation seem to be quite positive. Overarching statements have also made marking simpler.

The new scale seems to have eliminated some areas of doubt which previously existed in the minds of examiners such as the nature and degree of underlength script penalization, poor topic exposition and the extent to which credit should be given for aspects of lexical resource in the face of poor grammar. The scale also facilitates more precise assessments of punctuation and paragraphing. Word count rules seem considerably fairer although the need to count script totals was not widely welcomed by examiners.

Feedback from Test Administrators

From the Test Administrator perspective, the introduction of the new scale appears to have been relatively smooth. There was a fairly widespread perception that sufficient time had been given to retraining examiners. Several centres were favourably disposed to the new administrative procedures although not all centres echoed this sentiment.

One administrative change relates to the input requirements for ESOLCOMMS (Cambridge ESOL's administrative system used for the processing of IELTS). All eight criteria (compared to the original single entry) now require keying. The problems associated with entering additional scores have been widely acknowledged and in the majority of cases managed both quickly and efficiently. Nevertheless, centres observed that increased keying engendered significant extra workload for clerical administration staff. An increase in data processing time has added to the workload. However, the need for manually computed overall writing band scores has been removed which was previously a potential source of manual error in the rating process.

During the first six months of operation, an increase in instances of jagged profiles (where a candidate has achieved significantly different results across the four IELTS Modules) was reported by Test Centres. Some centres have also received a number of complaints regarding the writing scores since implementing the new writing assessment scales. Repeat candidates, it is claimed, continue to compare their previous writing scores with the new ones. Where the score achieved under revised conditions is lower this can be attributed to the additional assessment criterion introduced as part of the Revision.

According to administrators, incorrect completion of task penalty boxes or failure to complete score boxes correctly are relatively widespread. Hoverver, with increased monitoring at centre level, aspects of clerical administration will be checked more thoroughly.

Future areas of research

Future research will continue to be required in a number of principal areas. One area is the assessment of writing by electronic means. This does not mean that automatic electronic marking will take place. What it means is that the development of sophisticated scanning techniques will soon allow scripts to be sent from an examination centre back to Cambridge ESOL immediately after the completion of the test (*Electronic Script Management*). This development has a number of advantages. First of all it will allow for the use of a second marker if required (e.g. when a jagged profile occurs) assessing the script in question without any time delay so that the results of the test will not be delayed. Eventually, in cases where centres have the capacity to test candidates but where there is no certificated assessor, scripts will be marked by raters based elsewhere, e.g. the UK, without any delay to the candidate receiving results in the normal time-span.

A second area of research is the development and expansion of the PSN Extranet which helps administrators and examiners to keep up to date with developments and which provides support in terms of guidance and explanation. This use of technological developments, once given a chance to bed in, should enhance the general efficiency of IELTS (see Mitchell Crow & Hubbard 2006).

A third area of development concerns investigations into the use of word-processed text on rater behaviour (as opposed to hand-written text). The computer delivery of IELTS (CB IELTS) means that word-processed text for the Writing Module is becoming more common. A small-scale, preliminary study has already been undertaken with a group of examiners investigating the impact of word-processed text on rater behaviour (Shaw 2005c). The study took the form of a semi-structured, facilitated discussion in which raters were shown a range of scripts in handwritten and typed format. The purpose of the discussion with the raters was to gather soft feedback in order to ascertain the extent to which raters make judgments about responses in different formats. As this phase of the work was designed to be exploratory, any conclusions must be seen as tentative. The findings will, however, enable Cambridge ESOL to develop a more focused research perspective for future work as well as providing insightful observations into the way examiners might rate handwritten and word-processed text.

In addition, as part of the third area for investigation, the effect that training has on reducing the 'presentation' effect (typed or handwritten) needs to be explored further by replicating, on a larger sample and on larger groups of raters, the work of Russell and Tao (2004) who speculate that computer-printed text makes mechanical errors such as spelling and punctuation more visible and adversely affects rater scores. If subsequent trials offer evidence for the eradication of the presentation effect then a major barrier to providing test takers and students with the option of writing responses to composition-type questions may be removed.

A further area for consideration is the need to continue the trend set in the *Implementation Phase* – Phase 4, when every current IELTS examiner was trained, and either certificated or re-certificated as an IELTS writing examiner. It is clear that regular training, re-training and standardisation ensure that reliability standards are maintained. Once again, the use of web-enhanced training and standardisation will need to be investigated and developed to supplement such training, especially in the processes of refresher training.

One area, not mentioned above, is the response of candidates to the former and new versions of the Academic rubric. If resources can be found it would be worthwhile developing the trials undertaken by Bridges and Shaw (2004b and 2004c) by increasing the original number of candidates in order to investigate further their reactions to the two versions of the amended rubric.

Conclusion

One of the achievements of the IELTS Writing Revision Project is the sheer size of the project in terms of the commitment of human and financial resources in the areas of research studies, development, meetings, observations, trials and iterative processes that were felt necessary in order to accomplish the changes that were required. It is estimated that hundreds of people were involved in the project and thousands of hours committed to it. We have documented this Project to demonstrate the amount of attention given to the project in order to get it as right as possible before the *Operational* Phase commenced.

A comprehensive web-based research report (by Shaw and Falvey) documenting the entire revision project will be made available later this year. A full list of Cambridge ESOL's *Research Notes* articles relating to the IELTS Writing Revision Project can be found on: www.cambridgeesol.org/researchnotes

References and further reading

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

—(2004b) *Academic Writing Task 1 Rubrics Trial 26 March 2003, APU, ESOL*, Cambridge: Cambridge ESOL internal report 560.

—(2004c) *Proposed changes to the IELTS Academic Writing Rubrics: stakeholder perspectives*, Cambridge: Cambridge ESOL internal report 563.

Falvey, P and Shaw, S D (2006) IELTS Writing: revising assessment criteria and scales (Phase 5), *Research Notes* 23, 7–12.

Mitchell Crow, C and Hubbard, C (2006) ESOL Professional Support Network Extranet, *Research Notes* 23, 6–7.

Russell, M and Tao, W (2004) The Influence of Computer-Print on Rater Scores, *Practical Assessment, Research and Evaluation*, 9 (10), 1–14.

Shaw, S D (2002a) IELTS Writing: revising assessment criteria and scales (Phase 1), *Research Notes* 9, 16–18.

—(2002b) IELTS Writing: revising assessment criteria and scales (Phase 2), *Research Notes* 10, 10–13.

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

—(2003) IELTS Writing Revision Project (Phase 3): Validating the revised rating scale – a quantitative analysis, Cambridge: Cambridge ESOL internal report 513.

—(2003a) IELTS Writing Revision Project (Phase 3): Validating the revised rating scale – a qualitative analysis, Cambridge: Cambridge ESOL internal report 514.

—(2004a) Revising the IELTS Writing Rating Scale: senior examiner/trainer 're-training' feedback, Cambridge: Cambridge ESOL internal report 598.

—(2004b) Revising the IELTS Writing Rating Scale: feedback – trainer training and examiner training sessions, Cambridge: Cambridge ESOL internal report 608.

—(2004c) IELTS Writing: revising assessment criteria and scales (Concluding Phase 2), *Research Notes* 15, 9–11.

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

—(2005a) IELTS Writing Assessment Revision Working Group: Summary of Progress. June 2001 – May 2005, Cambridge: Cambridge ESOL internal report.

—(2005b) The IELTS Writing Assessment Revision Project: Operational Phase 5: APU Trial – April 2005, Cambridge: Cambridge ESOL internal report 671.

—(2005c) The impact of word processed text on rater behaviour: a review of the literature, Cambridge: Cambridge ESOL internal report 670.

TKT – a year on

NADEŽDA NOVAKOVIĆ, RESEARCH AND VALIDATION GROUP

Introduction

A year ago, Cambridge ESOL launched the Teaching Knowledge Test (TKT), a new addition to its range of teaching awards. The test has been developed for pre-service or practising teachers, and consists of three free-standing modules, each relating to a different aspect of knowledge about the teaching of English to speakers of other languages.

To date, the test has been administered at 66 centres in 24 countries around the world. Each module has been taken by more than 1000 candidates, bringing the total number of TKT test takers to around 4000. Seventy-seven percent of candidates took more than one module.

At every examination session, candidates were asked to fill in a Candidate Information Sheet (CIS), which provides, among other things, information on their age, gender, country of origin, teaching qualification and experience, their level of English language competence and reasons for taking the test. This article takes a brief look at the profile and the performance of the first generation of TKT candidates.[1]

[1] If candidates took more than one module, only one CIS form per candidate has been taken into account during the analysis of candidates' profiles.

Candidates' profile

Background

In TKT's first year, the majority of candidates came from Latin America, East Asia and Europe. Most of the test takers were female (82%). Eighty-seven percent of the candidates were between 26 and 50 years of age, with the majority belonging to the 31–40 age group.

Qualifications and experience

On CIS forms, candidates provided answers to two questions relating to their teaching qualifications and experience: whether they were qualified to teach English and other subjects, and how many years of teaching experience they had. Although 79% of candidates responded to being qualified to teach English in their country, the extent of their teaching experience ranged from no experience to 11 years or more of teaching experience. At the time of taking the test, most of the candidates had been in the teaching profession for 6 years or more (52%), while 28% have been teaching between 2 and 5 years. Twenty percent of the candidates had less than a year's experience (see Figure 1).

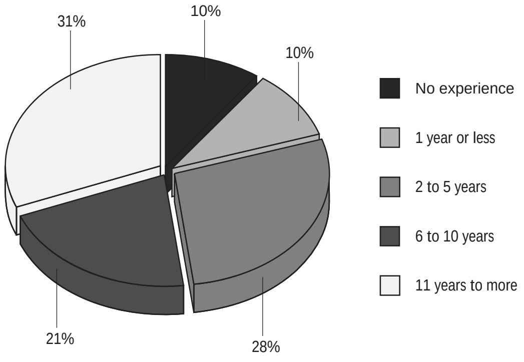

Figure 1 : Teaching experience of TKT candidates

The fact that candidates had a wide range of teaching experiences and came from many different teaching contexts, shows that both experienced and inexperienced teachers are finding the test relevant and useful. This is not surprising, as the test has been developed with both the novice and the experienced teacher in mind, suitable to be taken at any stage in a teacher's career. For less experienced candidates, TKT offers a step in their professional development and enables them to move onto higher-level teaching qualifications. For more experienced candidates, TKT offers an opportunity to refresh their teaching knowledge, and where applicable, forms a benchmark for teachers of subjects other than English who started teaching English as a foreign language.

English language proficiency

One of the questions on the CIS form asked candidates to rate their English language ability using a range of options, from elementary to advanced. Although TKT has not been designed to test candidates' English language ability, they are expected to have a minimum language level of B1 on the CEFR scale. They may, however, have a language competence well above this or they may be native speakers. Data collected so far show that 76% of candidates rated their English ability as either high intermediate or advanced, 20% as intermediate, while only 4% of the entire TKT candidature consider themselves as elementary or low intermediate English language speakers (Figure 2).

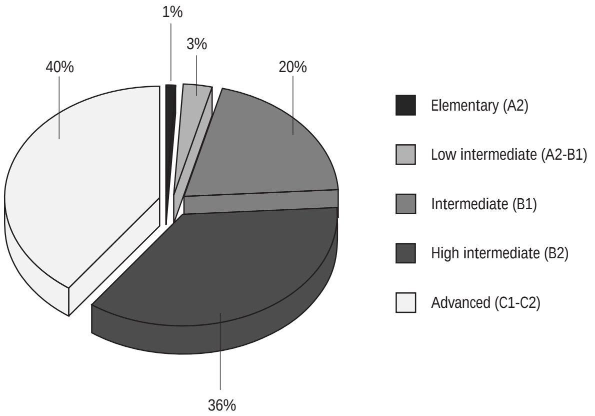

Figure 2 : Self-assessed English ability of TKT candidates

By collecting information on candidates' level of English, via self-assessment on CIS forms or by administering a short language test alongside TKT, Cambridge ESOL is able to monitor the language proficiency of TKT candidature over time and try to determine the extent to which it may affect candidates' performance on the test. This forms part of Cambridge ESOL's long-standing commitment to ensure that its exams meet the highest standards of validity; in this case, ensuring that candidates are assessed on their knowledge about teaching English, rather than their English language ability.

Reasons for taking TKT

The reasons that test takers gave for taking TKT varied from professional development and career advancement, to international recognition and employer's requirements. Professional development, however, was chosen as the main reason by the majority of candidates.

Candidates' performance

The previous section concentrated on the candidates' profile, showing that they came from a variety of backgrounds, had varied teaching experiences and different levels of English language competence. In this section, we take a brief look at candidate performance, focusing on the relationship between their level of teaching experience and the scores achieved on the test.

A detailed analysis of candidates' performance with respect to their teaching experience showed that, at times, there was indeed a statistically significant difference in performance between candidates with more and candidates with less teaching experience. A one-way ANOVA and a post-hoc test (Tukey) confirmed that, for example, there was a statistically significant difference on all Modules between the performance of candidates with no teaching experience and candidates with 11 years of experience and more.[2]

However, despite differences revealed by statistical analyses, the majority of candidates achieved the higher end of the scale, i.e. Band 3 and 4, with only a small percentage being awarded Band 1 or 2. Looking at Band 4, though, we find that it was awarded to 48% of candidates with 6 years of experience or more, 39% of candidates with 2–5 years of experience, and 22% of candidates with up to 1 year of experience. This would suggest that the more experienced candidates are, the more of them achieve Band 4.

The high percentage of candidates who achieved Bands 3 and 4 is not surprising, bearing in mind the nature of the test itself. As TKT has not been designed to test either candidates' language ability or their performance in the classroom, it is expected that any candidate who has covered the syllabus and is familiar with various approaches to teaching and learning, as well as relevant ELT terminology, should be able to perform successfully on the exam.

Conclusion

In its commitment to produce examinations which meet the highest standards of test validity, Cambridge ESOL continues to monitor TKT candidature, noting any changes in the test user population that may lead to further test developments and revisions. With this aim in mind, it engages in research and validation activities that focus on test taker characteristics, and their interaction with and effect on the validity of the test. Future research plans with respect to TKT include the comparison of performance of those candidates with specialist and non-specialist knowledge of English language teaching.

[2] One must, however, bear in mind the relatively small number of the least experienced candidates compared to the number of candidates with (relatively) substantial teaching experience (see Figure 1).

IELTS Joint-funded Research Program Round 12: call for proposals

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

As part of their ongoing commitment to IELTS-related validation and research, IELTS Australia and the British Council are once again making available funding for research projects in 2006/7. For several years now the partners have issued a joint call for research proposals that reflect current concerns and issues relating to the IELTS test in the international context. A full list of funded research studies conducted between 1995 and 2001 appeared in *Research Notes* 8 (May 2002); studies conducted between 2002 and 2004 appeared in *Research Notes* 20 (May 2005), and *Research Notes* 23 (February 2006) contains a list of studies funded in 2005.

Such research makes an important contribution to the monitoring and test development process for IELTS; it also helps IELTS stakeholders (e.g. English language professionals and teachers) to develop a greater understanding of the test.

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

Details of the call for proposals including application forms, timescale and guidance on topics and resources can be found on the IELTS website under Grants and Awards: http://www.ielts.org