Research Notes

Contents

Editorial Notes

Welcome to issue 37 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The topic of this issue is the Common European Framework of Reference for Languages (CEFR) and its impact on language assessment, specifically how it is used in Cambridge ESOL. This issue opens with an introduction by Cambridge ESOL's Chief Executive, Dr Michael Milanovic, who has been involved with the CEFR since its inception and who outlines his own stance on the CEFR and describes its influence on his own work and that of Cambridge ESOL. This is followed by a series of articles by Cambridge ESOL staff which describe in more detail the practical, theoretical and wider issues that we face on a daily basis in relation to the CEFR.

Neil Jones provides his own perspective on the use of standard setting when constructing a multilingual proficiency framework. He explores issues related to absolute and comparative judgement and discusses in detail a ranking approach to align different languages and tests to a common scale as a pre-cursor to standard setting. Next, Hanan Khalifa and Angela ffrench share Cambridge ESOL's approach towards relating examinations to the CEFR. They provide examples of how the CEFR *Manual* procedures can be embedded effectively within existing systems and processes.

The issue then moves to provide cases studies of using the CEFR and its toolkit. Szilvia Papp and Angeliki Salamoura consider whether young learner examinations can be aligned to the CEFR, bearing in mind that the CEFR was not intended to be a framework of reference for children's language learning, teaching and assessment. They investigate the extent to which the CEFR can offer assistance to practitioners within a Young Learners context and report on an exploratory study that made use of the *Manual's* recommendations for alignment.

When relating examinations to the CEFR, test developers need to use exemplar materials. Evelina Galaczi and Hanan Khalifa describe the production of Cambridge ESOL's new DVD of oral performances which contains a set of speaking test performances that exemplify a range of CEFR levels. Their article is of use to readers who may want to compile their own set of exemplar materials.

Next, Hugh Bateman points out that alignment to international standards should consider aspects of context validity. He focuses on CEFR B2 level and teases out some salient features of the level in terms of the linguistic demands of a writing test within an LSP context. On a similar topic, Angeliki Salamoura and Nick Saville, referring to one of the aims of the English Profile Programme, namely, the identification of criterial features for each CEFR level, provide some criterial features of English across the CEFR levels.

We finish this issue with a bibliography containing all of the references from this issue on CEFR and Cambridge ESOL. This is available as an offprint on our website, along with the other articles in this issue, at: www.CambridgeESOL.org/rs_notes

Editorial team for Issue 37: Nick Saville, Hanan Khalifa, Fiona Barker, and Caroline Warren.

Cambridge ESOL and the CEFR

MICHAEL MILANOVIC CEO, CAMBRIDGE ESOL

Introduction

I am very pleased to be able to introduce this special issue of *Research Notes* which provides a number of perspectives on the alignment of examinations to the Common European Framework of Reference (CEFR), and to be able to outline my own stance on this issue.

I would like to start by saying that I fully support the concept of the CEFR as a widely-used, common framework of reference based on six broad reference levels and an 'action-oriented' approach to language teaching and learning. The CEFR has certainly helped to raise awareness of language issues and has provided a useful focus for researchers, policy makers, assessment providers, teachers and so on. I also fully endorse the particular approach to innovation within language policy and education practised by the Council of Europe.

However, I have a concern about some uses and interpretations of the CEFR, particularly recent applications which are in my view inappropriately prescriptive or regulatory in nature. I will explain why I think this is problematical below. Let me begin with some background to these issues and others by outlining my own involvement with the CEFR dating back to the early 1990s.

Background to Cambridge ESOL's involvement with the CEFR

I participated in the *Rüschlikon Intergovernmental Symposium* in November 1991 (see Council of Europe 2001a:5), and was invited to join the Advisory Group which commented on the development of the CEFR throughout the 1990s. I worked with and commented on the first drafts of the CEFR which were produced between 1995 and 1998. In March 1999 I hosted a meeting in Cambridge which included two of the authoring group and several other stakeholders to discuss the timescales and tasks needed to revise the manuscript in time for publication in 2001 – the first European Year of Languages. This led to the final editing coordinated by John Trim in 2000. In light of my close involvement with the principles and practices of the CEFR, I have set out the agenda for Cambridge ESOL's approach to framework development and have also played a leading role in ALTE's work in creating its multilingual system of proficiency levels. Both projects were initiated in the early 1990s and then developed at the same time that the CEFR was being written and discussed in the mid-1990s. Both have evolved over an extended period in what North has called 'a process of convergence' (North 2008).

In line with the original recommendations of the Advisory Group, I have encouraged the development of the so-called 'toolkit' to allow users to make better use of the CEFR for their own purposes, and have overseen or been directly involved with a number of such initiatives, for example:

1. Coordinating the development of a *Users' Guide for Examiners* in 1996 (published by Council of Europe) – now under revision by ALTE as a *Manual for Language Test Development and Examining*.
2. Developing the EAQUALS/ALTE European Language Portfolio (ELP) both in hardcopy and electronic forms (from 2000).
3. Providing support for the authoring and piloting of the draft *Manual* for aligning examinations (since 2002/3).
4. Contributing to benchmarking materials and examples of test items to accompany the CEFR (from around 2004).
5. Developing content analysis grids for speaking and writing materials (based on ALTE projects dating back to 1992).
6. Hosting of a 'case study' event to report on the piloting of the draft *Manual* (Cambridge, December 2007) and publication of the proceedings in the Studies in Language Testing series (Martyniuk forthcoming).
7. Developing Reference Level Descriptions for English – The English Profile Programme (since 2005).

Cambridge ESOL's involvement with the Council of Europe has an even longer history dating back to 1980 when the concept of a multi-level system of Cambridge examinations began to emerge in light of Wilkins' work on proficiency levels (see Trim 1978) and starting with the addition of a *Threshold Level* test (PET) to the well-established FCE and CPE examinations. In 1990 the revised *Waystage* and *Threshold* specifications (which had been partly sponsored by Cambridge) formed the basis of the test specifications for the new KET and updated PET and further additions and revisions to existing examinations saw the process of convergence taking place to achieve this goal, as noted by North (2008:31–2).

Cambridge ESOL's common scale

One of the first things I set out to achieve when I joined Cambridge in 1989 was to ensure that Cambridge ESOL developed the concept of an *empirically-derived common scale* that allowed for the systematic ordering of examinations according to level and purpose of use (see the series editor's note in *Studies in Language Testing* volume 1, Bachman, Davidson, Ryan and Choi 1995:vii–xi).

The empirical underpinning for the system was achieved by introducing an item banking approach which allows for all examinations to be calibrated on a common scale within a psychometric framework. Since the inception of the common scale many millions of candidates at all levels have taken the Cambridge examinations and their responses have allowed the scale to be incrementally refined based on analysis of these data within the framework. (See North and Jones' 2009 paper for the

Council of Europe to accompany the revised *Manual*; also Maris 2009 for discussion of test equating using IRT in the context of standard setting).

The work of ALTE and multilingual aspects of standard setting

In the early 1990s ALTE also began to develop its own multilingual level-system based on a systematic analysis and comparison of test content across languages and the development and calibration of its own *Can Do statements* (represented as Appendix D in the CEFR, 2001:22 and 244–57). The ongoing development of the ALTE framework and the challenge of cross-language alignment of examinations within such a framework remained a concern. This *multilingual aspect of standard setting* became a preoccupation of the ALTE Special Interest Group which has worked with the pilot-version of the *Manual* since about 2004. Although not discussed at length in the *Manual*, the topic of multilingual alignment was taken up by North and Jones in their 2009 paper, and by Jones at the standard setting seminar prior to the EALTA conference in Athens in 2008. His paper has recently been published in the proceedings of the event (see Figueras and Noijons 2009) and a version of his paper also appears in this issue of *Research Notes* (see Jones 2009). At a European level this challenge has recently been picked up by the SurveyLang group led by Cambridge ESOL to deliver the *European Indicator of Language Competences*, initially in five languages (see www.surveylang.org, also Jones and Saville 2009).

In light of these observations, I would now like to focus on three main points which underpin my own stance and which have influenced the work of Cambridge ESOL over nearly 20 years. This stance has led to a coherent and well-documented approach which is reiterated in the papers in this edition of *Research Notes* (see the bibliography at the end of this issue for earlier coverage of this topic).

Three points to remember about the CEFR

Point 1

First, I think we should constantly remind ourselves that the CEFR itself is deliberately underspecified and incomplete. It is this feature which makes it an appropriate tool for comparison of practices *across many different contexts* in Europe and beyond. On the one hand it is useful as a *common framework* with six broad reference levels, but on the other it is not applicable to all contexts without user intervention in order to *adapt it to suit local purposes*.

The three main authors of the CEFR, Coste, North, and Trim, made this point very clearly in the text itself and they have all repeated it on numerous occasions in subsequent presentations on the framework and its principles. So, for example, in the *introductory notes for the user*, the following statement is emphatically made: 'We have NOT set out to tell practitioners what to do or how to do it' (Council of Europe 2001:xi). This is reiterated throughout the text by the use of the phrase: 'Users of the framework may wish to consider and where appropriate state…' (e.g. Council of Europe 2001:40).

Subsequent work on the 'toolkit' has also followed this lead. For example, the authors of the *Manual* for aligning language examinations to the CEFR stress this point when they state that the *Manual* '… is not the sole guide to linking a test to the CEFR and there is no compulsion on any institution to undertake such linking' (Council of Europe 2009:1).

This was also the approach which I adopted in 1995/6 when, on behalf of the Council of Europe, I coordinated the drafting of a *Users' Guide for Examiners* to accompany the CEFR. By design, this was a non-prescriptive document which sought to highlight the main principles and approaches to test development and assessment which users could refer to in developing tests *within their own contexts of use*, and not a cook book for developing test questions based on the illustrative scales.

More recently in his plenary paper presented at the Council of Europe Policy Forum on the use of the CEFR (Coste 2007), Coste has described how contextual uses which are seen as deliberate interventions in a given environment can take 'various forms, apply on different levels, have different aims, and involve different types of player'. In his view:

> 'All of these many contextual applications are legitimate and meaningful but, just as the Framework itself offers a range of (as it were) built-in options, so some of the *contextual applications* exploit it more fully, while others extend or transcend it.'

When considering alignment questions, this fundamental principle must be the starting point and constantly borne in mind because there are important implications which follow on from this. For example, it is important to remember that the CEFR is not intended to be used prescriptively and that there can be no single 'best' way to account for the alignment of an examination within its own context and purpose of use. As Jones and Saville (2009:54–5) point out:

> ' … some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring a context to the CEFR. The transitivity is the other way round. The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context which determines the final meaning of the claim. By engaging with the process in this way we put the CEFR in its correct place as a point of reference, and also contribute to its future evolution.'

A particular concern of mine relates to the status of the 'illustrative scales of descriptors' as they are called, and their recent uses in overly prescriptive ways (i.e. against the intentions of the authors) particularly in the context of standard setting. In one of the pre-publication drafts of the Framework which I worked with, entitled *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference* (Council of Europe 1998), these scales were included in the appendix as examples and did not occur in the body of the text. The only scales to be included in the main text were the common reference levels (later to become Tables one, two and three in the published version, Council of Europe 2001a:24–9).

This layout of the text visibly reinforced the different status and function of the general *reference levels* and more specific *illustrative scales*. This was an approach which I favoured personally since it underlined the very tentative nature of the illustrative scales, many of which were uncalibrated and indeed were under represented, particularly at the C-Levels. Given the vigour with which some people have recently attempted precise alignment using these scales, despite their obvious and clearly stated deficiencies, I feel justified in my original view that it would be dangerous to give the illustrative scales too much prominence.

In Chapter 8 of the 1998 version which was entitled 'Scaling and Levels' the tentative status of the illustrative scales was made clear in the following paragraph:

> 'The establishment of a set of common reference points in no way limits how different sectors in different pedagogic cultures may choose to organise or describe their system of levels and modules. It is also *to be expected that the precise formulation of the set of common reference points, the wording of the descriptors, will develop over time* as the experience of member states and of institutions with related expertise is incorporated into the description.' (Council of Europe 1998:131; emphasis added)

Since the publication of the CEFR in its finalised form in 2001, the second point in this paragraph which emphasises the tentative nature of the illustrative scales has tended to be forgotten or at least downplayed by some users. This may be due in part to the way that the final text was edited.

Many of the less well validated illustrative scales remained in the final text, but for pragmatic reasons the authoring group decided to incorporate them into the main text rather than keep them in the appendix. Four appendices were used to illustrate several projects involving the development of scale descriptors; Appendix B (Council of Europe 2001:217) was used to describe the development of the 'illustrative scale descriptors' which was part of the Swiss research project conducted by North (later published as a book based on his PhD; North 2000).

But the points made by the authors in 1998 still remain true; in other words, the functional and linguistic scales were there to illustrate the nature of the levels rather than to define them precisely. While some of the scales might prove stable across different contexts, there should not be an expectation that they all will. This has important implications for the use of the 'illustrative scales of descriptors' in alignment procedures; for example, given their status, individual scales should only be used with great care in any kind of standard setting exercise. Indeed it is hard to see how, over and above a very general approximation to the levels, standard setting using the current scales can be considered a satisfactory procedure.

North himself (2007b) notes that the 'fluency' scale was useful in linking the *ALTE Can Do project* to the framework (based on values from the Swiss project he had carried out) but that other scales were not robustly calibrated, and there were significant gaps at the A1 and C levels (see North's presentation made at the 23rd ALTE conference, Sèvres, April 2007 – available from ALTE website: www.alte.org).

Somewhere along the way, these very real concerns of a principal author of the scales have been lost. Indeed, given the origins and status of the scales it is perhaps unfortunate that there has been a somewhat one-sided reading of its text, as noted by Coste (2007), another CEFR author: 'In various settings and on various levels of discourse … people who talk about the Framework are actually referring only to its scales of proficiency and their descriptors.'

For a summary of Trim's views on the framework which follow similar lines see Saville 2005 (*An interview with John Trim at 80*).

Point 2

This leads to my second point which I think is even more significant. If the CEFR is to have lasting impact, then its principles and practices should be integrated into the routine assessment procedures of an examination provider so that alignment arguments can be built up over time as the professional systems develop to support the claims being made. This entails working with the text of the CEFR as a whole and adapting it where necessary to suit specific contexts and applications. In my view, it is unlikely that any single report can provide satisfactory evidence of alignment. On the contrary, a single standard setting exercise should not be taken as sufficient evidence and examination providers should seek to provide multiple sources of evidence accrued over time.

Standard setting events which are conducted as one-off procedures, do not provide enough evidence for consistent interpretation of any level system. If necessary, alignment arguments should remain tentative and be modified later in light of additional evidence when it becomes available. This should be expected rather than be seen as a problem.

Point 3

My third point also relates to this. When we talk about assessment, then alignment arguments and assessment standards need to be maintained in the long-term using a range of techniques and professional processes, including:

- Item banking to establish common measurement scales and to allow for both item-based and person-based equating to be employed in test construction and in the monitoring of standards over time.
- Routine test validation processes to quality assure test outcomes.
- Iterative cycles of test development and revision.

More specifically this means that the recommendations found in the *Manual* on how to use the CEFR and other resources supplied by the Council of Europe for alignment purposes (e.g. familiarisation activities with stakeholders and standard setting exercises of different types whether task-based or person-based), need to be integrated within the standard procedures of the assessment provider and should not be seen as 'one-off events'. This is particularly true for an examination board like Cambridge ESOL which works with (literally) thousands of stakeholders in developing, administering, marking and validating many different types of examination within a consistent but evolving frame of reference. For example, in 2010 over 400 administrations of different Cambridge examinations will take place, all of which include the assessment of four skills (including face-to-face speaking tests). Given the complexity of this operation, the arguments for alignment to external reference points need to be developed on a case-by-case basis and must be one part of the broader validity argument which is needed to support the appropriate uses of each examination.

Linguistic competences and the application of the CEFR to English

Finally I would like to return to the underspecification of the CEFR and to consider what this means for relating particular language examinations to the framework. The CEFR is neutral with respect to language and, as the *common* framework, must by necessity be underspecified for all languages. This means that specialists in the teaching or assessment of a given language (e.g. Cambridge ESOL for English) need to determine the linguistic features which increasing proficiency in the language entails (i.e. the user/learner's competences described in Chapter 5 of the CEFR). Such features are peculiar to each language and so the CEFR must be adapted to accommodate the language in question.

Cambridge ESOL's testing system has developed alongside the CEFR and has "converged" with it over the past two decades; it is now able to provide rich data and analysis to help refine the CEFR as applied to English. This is an important role for a responsible organisation to fulfil and very much in keeping with the original intentions of the Council of Europe. Our aim is to facilitate understanding and collaborative activities rather than to regulate or dictate to others what they should or should not do. An example of this in practice is the *English Profile (EP) Programme* (see www.englishprofile.org; also *Research Notes* 33).

A major objective of English Profile is to analyse learner language to throw more light on what learners of English *can* and *can't do* at different CEFR levels, and to address *how well* they perform using the linguistic exponents of the language at their disposal (i.e. using the grammar and lexis of English). One of the main inputs to this analysis is provided by the Cambridge Learner Corpus which contains 35 million words of learners' written English from levels A2 to C2 of the CEFR. The researchers are already providing evidence of 'criterial features' of English which are typically found in the writing of learners at the different CEFR levels (see Salamoura and Saville 2009 in this issue). Of course this data alone does not provide an adequate sample and so part of the EP Programme includes the collection of additional data from learners within the 'EP Network', including more written data and also focusing on spoken English as well.

We are now in a position to begin a systematic and empirically-based approach to specifying more precisely how the CEFR can be operationalised for English, and this in turn will lead to better and more comprehensive illustrative descriptors (particularly at the bottom and top of the scale). In this way the CEFR will become the really useful tool that it was intended to be.

Conclusion

In conclusion, I would like to reiterate my support for the principles and practices of the CEFR and for what I see as the main strength of the CEFR so far, its use as a communication tool. Within the common framework of levels, Cambridge ESOL has attempted to make the interpretation of examination results as transparent and meaningful as possible and the development of functional descriptors (Can Dos) has been useful in promoting better communication between stakeholders.

But as I have noted above, I think it is also important to draw attention to some limitations and uses of the framework for which the CEFR was not designed. Some of these limitations were acknowledged by the original authors and some others have also been noted in the literature over the past few years (see for example: Alderson 2007, Fulcher 2004, McNamara and Roever 2006, Weir 2005a). In particular the uses of the CEFR which seek to direct or control users should be resisted.

As a responsible assessment provider, Cambridge ESOL also seeks to provide leadership in the field of language testing, and I feel that it is important for Cambridge ESOL to address these issues explicitly. That is why I have attempted to make my own stance very clear. By working collaboratively with the CEFR, the shortcomings of the illustrative scales and linguistic content can be addressed more effectively, with data being collected to enable well-informed refinements to be made as our understanding increases. As Lynda Taylor has concluded in an earlier article for *Research Notes* (2004:3):

> 'As we grow in our understanding of the relationship between IELTS, other Cambridge ESOL examinations and the CEFR levels, so the frame of reference may need to be revised accordingly.'

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting

NEIL JONES RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Acknowledgement

This is an updated version of a paper in the proceedings volume from the pre-conference colloquium on *Standard Setting Research and its Relevance to the CEFR* which took place on 6th May 2008 at the EALTA conference in Athens. Our thanks go to the publishers of that volume – Cito (the Institute for Educational Measurement in the Netherlands) and EALTA (European Association for Language Testing and Assessment) – for permission to reproduce it here.

Standard setting and the CEFR: the problem

For some years now I have been expressing an essentially sceptical view of standard setting for the purpose of constructing a multilingual proficiency framework – that is, as advocated in relation to the CEFR. My position dates from 2003, when I started work on the construction of a similar framework for a new set of UK qualifications, called *Asset Languages*. This project was offered as a case study of applying the pilot versions of the *Manual* for aligning exams to the CEFR (Council of Europe 2003a). When it came to presenting the case study I had to confess that it was more about where we had not followed the pilot version of the *Manual* than about where we had followed it. I also emphasised that this was not in itself a judgment on the *Manual*, but rather reflected the development procedures that we had been forced to adopt by the multilingual scope (twenty-five languages) and tight schedule of that particular project. What the Asset Languages project did impress on me was the need to look for ways of taking a high-level, top-down view of the process of constructing such a framework, with respect to its vertical dimension of progression through levels, and its horizontal dimension of alignment and comparability across languages. This requires us to find methods of working explicitly with these two dimensions, rather than dealing with each language and level separately in the optimistic belief that decisions made at micro-levels will lead to the emergence of a coherent whole.

Clearly, there are many different aspects to implementing a top-down model in a test development project and pursuing the (finally unachievable) goal of perfect comparability across languages. The *Manual* is a valuable resource, which undoubtedly facilitates the adoption of a top-down model, with the CEFR providing the coherent framework to which each language, and each group of language learners, may be aligned. (Incidentally, I prefer to speak of aligning language learners to the CEFR rather than language tests, because in the end the process concerns the validity of tests, and validity concerns the inferences we make about the learners who take them. If a test for young learners is difficult to align to the CEFR, it is because the argument that links young learners' language performance to the CEFR is currently harder to construct.) However, I was not fully convinced by the treatment of standard setting in the pilot version of the *Manual*, for the following reasons.

Firstly, I find the use of terminology slightly idiosyncratic. The term 'standard setting' is used in Chapter 5 to refer to task-centred approaches and to objective tests, while learner-centred approaches are treated as 'external validation' in Chapter 6. This suggestion of logical and temporal priority – that task-centred standard setting is an essential first step, and that learner-centred approaches are an option for some later validation stage – does not seem to reflect the treatment of these two broad approaches in the literature, where both are simply referred to as standard setting. Secondly and chiefly, I feel that the use of task-centred standard setting approaches in constructing a multilingual framework is a misapplication of techniques to a situation where their underlying premises and justification do not hold.

I can make this clearer by describing what I would call a *classical* standard setting context – the one in which many of these approaches developed – and contrasting it with our purpose in relation to the CEFR. In this way I can identify the issues which I believe require our particular attention.

Let's take as the classical context the professional licensure exam: say, for example, a (hypothetical) one hundred item multiple choice test for nurses. We can characterise this context in terms of the following premises:

1. The judges and candidates are members, or prospective members, of a specific professional community.
2. The test tasks relate to discrete items of professional knowledge.
3. The judges are qualified to say which items a practitioner should master.
4. Hence the notion of 'minimal competence' has substantive meaning.
5. The buck stops with the judges, who are responsible to the public. Judgements are not 'correct', only defensible in terms of balancing the interests of the candidate nurses and the public whom they will serve.
6. The frame of reference is the profession and its stakeholders, and no judgements have implications outside this frame of reference.

7. The judges' professional and cultural background (for their practice is culturally embedded) impacts on their decisions and actually reinforces their validity (within that culture).

The CEFR context clearly differs in several important respects. Listening and Reading are skills: tests do not simply measure discrete atoms of knowledge, but attempt to tap hidden mental processes (violating premises 2 and 3 above). Hence we are dealing with an indirectly observable continuum of ability: the notion of minimal competence, or any discrete level of competence, is hard to pin down (violating premise 4).

The frame of reference is languages across Europe, and so *all* judgements have implications which extend beyond the immediate context of a particular test or language (violating premise 6). Judgements *can* and *must* aspire to be 'correct' in the sense of *consistent* with other judgements being made within the wider frame of reference (violating premise 5). Therefore the culturally-determined nature of judgements, far from reinforcing their validity, becomes a serious threat to it (premise 7). This last point in particular presents the major challenge for aligning standards across languages. Clearly, the whole purpose of the CEFR is to provide that practical point of reference that enables a common understanding of levels. But level descriptors are not wholly concrete or definitive. They require interpretation, and our default expectation must be that different countries' interpretations will be culturally determined (in a broad sense) and therefore may differ. This is, of course, not just a hypothetical problem, but is a recognised current practical issue which is now beginning to be addressed, most notably in the important multilingual benchmarking event held by CIEP in Paris in June 2008.[1]

In this section I have argued that the assumptions or premises which justify orthodox task-centred standard setting approaches are violated in the case of linking language tests to the CEFR. It is necessary to look at the problem in a different way.

Absolute and comparative judgement: rating and ranking

If the CEFR's frame of reference takes in all European languages then clearly the correctness of a standard set for any language can only be evaluated by comparison with other languages. Instead of attempting absolute judgements about the level represented by a score on a Reading or Listening test, or a sample of performance in Writing or Speaking, we need to think in terms of comparative judgements: is this Reading task in language X harder or easier than this task in language Y? Is this sample of Speaking in language X better or worse than this sample in language Y? The basic act of judgement in a multilingual frame of reference is thus not rating, but ranking. This reflects a general principle that constructing a framework is logically a two-stage process: first we construct a common measurement scale, and second we set standards.

I can try and make this point more clearly by offering an analogy with measuring and interpreting temperature. Historically the first step was to construct a practical measuring instrument – a thermometer. The next step was to calibrate it – that is, put a numbered scale on it. It evidently made sense to devise a standard scale and ensure that all thermometers reported using it. Today Celsius has become the standard scale for most purposes. Only at this point did it become practical to develop interpretations of points on the scale. We have been able to develop and share a sophisticated understanding of how seriously to treat varying degrees of fever precisely because our measurement instruments are aligned to the same scale.

Standard setting research and its relevance to the CEFR

To relate this back to our multilingual framework: it makes logical sense *first* to align tests across languages to the same scale, and only then to develop interpretations – i.e. set standards. Those interpretations will then apply equally to all of the aligned languages. Of course, what makes logical sense is not always possible in practice – it certainly wasn't in the case of *Asset Languages*, and neither is it in the case of the CEFR, where so much has already taken place. However, what I propose here could contribute to the current iterative process of progressive approximation to the intended meaning of the CEFR levels.

By focusing on comparative judgements – ranking – we can achieve the alignment of language tests and performances to the same scale. We should find this an easier, more practical task because human beings are much better at making comparative judgements than absolute judgments. Bramley (2005) quotes Laming (2004), who goes so far as to say: 'There is no absolute judgment. All judgements are comparisons of one thing with another.' It also addresses the more fundamental question. In my understanding, the question 'Is my B1 your B1?' is first a question about equivalence, and only second a question about the meaning of B1.

And if we can answer this question we are already much better placed to answer the second question – the one about interpretation, or standards. Obviously, a comparative approach cannot remove the need for standard setting at some stage, but by placing it at a logically later stage – after the alignment of languages to a common scale – it dramatically reduces the scope of standard setting. The standard is set once but applies equally to all aligned languages. Subsequent languages can be aligned to the same framework by a relatively simple comparative exercise. There is no need – in fact it is not possible – to do standard setting separately for each such language, because the act of alignment applies the standard already set. Thus we can conclude that the logic of a multilingual framework is such as to severely constrain the freedom of judgements relating to individual languages. If we accept this then there follow further possible conclusions for the methodology of framework construction.

Concerning objectively-marked tests of Reading and Listening, it remains a problem for standard setting to establish meaningful cut-offs on what are essentially

[1] Centre international d'études pédagogiques, see www.ciep.fr

continuous measurement scales relating to indirectly observed mental processes. For these skills in particular it is comparability of measures which is paramount. If we can develop a measurement scale, or appeal to some existing one, which defines levels rationally in terms of the way they relate to substantive learning gains, likely learning hours between levels, or the definition of accessible learning targets, then we can argue that this scale could be applied by default across languages. As North (2006) suggests, the CEFR has developed out of a concept of levels which are appropriate for broad groups of learners at particular stages in their learning career, and taken together define a progression which makes sense as a 'learning ladder'. Taylor and Jones (2006) describe the development of the Cambridge ESOL levels and their relationship to the CEFR in similar terms.

This was the approach adopted with the Reading and Listening scales for Asset Languages, where we adopted as a prototype or template the common scale upon which the Cambridge ESOL levels have been calibrated. That is, experience of working with these scales, which of course depends on an item banking, IRT scaling methodology, gave us a useful expectation of how a scale for similarly tested skills should look. I've written about this idea elsewhere in relation to scaling the CEFR levels for objective tests (Jones 2005); it is mentioned here just to reinforce the point that in a multilingual framework freedom of standard setting judgement is very severely constrained, one of the constraints being the proportional placement of levels on a measurement scale developed using a particular kind of response data.

Data collection and analysis for a ranking approach

I have stated that ranking allows us to align languages to a common scale, which in turn allows us to set the same standards for all of the aligned languages. I will now look at methods that we can use.

Bramley (2005) reviews comparative approaches. The earliest of these is Thurstone's paired comparison method (Thurstone 1927), which is based on the idea that the further apart two objects are on a latent trait, the greater the probability of one of them 'winning' a comparison. Thus from a set of dichotomous judgements (e.g. of 'better' or 'worse') one can estimate not simply an ordinal ranking, but the relative location of each object on an interval latent trait scale. Thurstone's model can be implemented in different ways, of which the most computationally tractable is a Rasch formulation (Andrich 1978). However, a practical problem found by Bramley and others using paired comparisons is the repetition and sheer number of paired judgements required. A ranking approach, where more than two objects are compared, is thus an attractive alternative. One analysis approach is to decompose ranking data into paired comparisons, although because these are of necessity self-consistent they lack independence and thus violate Rasch model assumptions, exaggerating the length of the measurement scale. Alternatively rankings can be used as categories in a Rasch partial credit model. Here the top-ranking object 'scores' 1, the second 2 and so on, for each judge involved. Bramley (2005) shows that the methods produce highly correlated results. Linacre (2006) reviews different methods of analysing rank-ordered data.

Bramley (2005) treats the case of a National Curriculum test of Reading attainment for pupils aged 14, and of equating performance on test versions from one year to the next. He distinguishes standard setting from what he calls *standard maintaining*: a comparative approach is used here to attempt to apply the standard from a previous year to the current year. The objects of comparison were scripts containing pupils' responses to short-answer Reading questions. This was thus a comparison of pupils' performance. It allowed the pupils from the two years to be aligned on a single ability scale, from which equivalent cut-off scores for the two test versions were estimated by linear regression of marks on ability. Results from this ranking study were found to agree well with an equating based on different information.

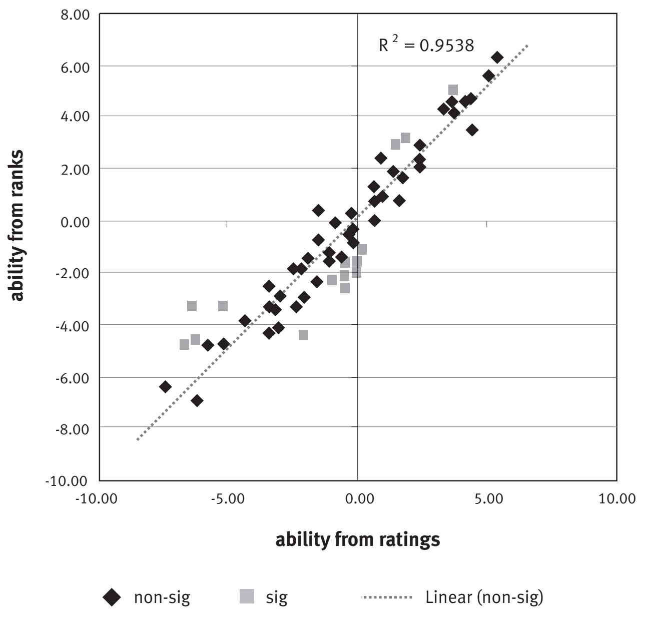
The multilingual benchmarking conference organised by CIEP at Sèvres in June 2008 also focused on performance, this time the skill of Speaking. Two kinds of data were collected. At the conference itself judges rated video performances against the CEFR, using a similar methodology to earlier such events conducted for French, German and Italian between 2004 and 2006, but with the difference that ratings were elicited in a 'cascade' design using English and French as 'anchor' languages: working in one group (on English and French), then in two and then three parallel subgroups, each dealing with three languages (i.e. English, French and one other).

Prior to the conference ranking data were collected from the same judges, using a specially-developed web-based platform which allowed them to view samples and record their ranking by dragging samples to re-order them in a list. The allocation of samples for the ranking exercise was such as to ensure that each judge rated in two languages, and that there was linkage in the data across all samples and languages.

Figure 1 compares the abilities estimated from rankings and ratings for the set of samples submitted to both procedures. The correlation is high. The lighter squares are outcomes which are significantly more discrepant than measurement error would allow. Clearly there are some significant differences in the outcomes, but given the fact that the ranking exercise took place before the conference, and was done individually online without guidance, discussion or familiarisation with the procedure, this is not surprising. This first study makes the ranking approach look very promising.

There remains the issue of whether a comparative approach can be made to work for the task-centred case of Reading or Listening as measured by objectively-marked items. This is the most difficult case – the one where, I have argued, the standard setting approach proposed in the *Manual* is least convincing. The focus here is not on samples of performance, but on comparison of test items across languages. Apart from a small workshop I conducted at the ALTE conference in Vilnius in November 2007 I have no data to base a claim on. However, I would be reasonably hopeful that a comparative approach can be made to work,

Figure 1: Ranking and rating compared (Sèvres June 2008)



because the direct comparison of items remains a simpler, more concrete cognitive process than those demanded by orthodox task-centred standard setting methods. Procedures could be explored in which judges would be observed ranking items by difficulty for a single language as well as across two languages. The cross-language comparison could be done in various ways, with knowledge of the relative difficulties of none, or one, or both of the item sets. Outcomes could be correlated with item calibrations from empirical data to derive indices of probable accuracy with respect to the single-language and by extension to the cross-language case. Thus we should be able to find ways of estimating standard errors of an alignment, interpretable in ways that the standard errors provided by standard setting are not.

Conclusion

I have proposed that a ranking approach offers a practical way to align different languages and tests to a common scale, and that it is logical to do this as a separate step prior to standard setting. This priority reflects our fundamental concerns in building a multilingual proficiency framework: first, to establish equivalence of levels, secondly to assign meanings to them. This priority may not be obvious to the majority of groups concerned with aligning a test or a set of exam levels to the CEFR, because they are interested in a single language. However, as we begin to establish a core of credibly-linked languages, with the associated benchmark exemplars of performance, and calibrated sets of tasks for objective skills, both the feasibility and the compelling arguments in favour of adopting a comparative approach will become clearer.

There is further work to do developing and validating comparative methodologies for aligning performance skills and objective test items, and deriving indices enabling us to evaluate outcomes. There are undoubtedly wider and more fundamental issues concerning the nature of comparability, which are highlighted by the approach but already lie at the very heart of a frame of reference like the CEFR. On what basis *can* younger learners be compared with adults, or formal learning compared with informal acquisition through immersion? These are not questions to address here, but from observing benchmarking events it seems that certain conventional ground rules have to be agreed. Perhaps the nature of these rules still requires better articulation and theoretical justification. My position here is simply that if we accept the utility and practicality of aligning different learning contexts to the CEFR then the comparative approach is a good way of doing so.

In this paper I have stressed the priority of scale construction. It occurred to me during the Athens colloquium that some of the cases presented, even though their focus might be on a single language and even a single level, were indeed using standard setting methods to address what were properly scaling issues. Having a valid and reliable approach to test construction is a pre-requisite for addressing standards, because without it the standard may simply fluctuate from session to session. The *Manual* in its final form makes this point, but in my opinion still underplays the importance of developing an item-banking, IRT-based approach to test construction that can ensure consistency of standards across sessions. Thus it gives the impression that alignment to the CEFR might be satisfactorily achieved by a one-off standard setting event.

Also, the *Manual* in its final form (Council of Europe 2009) has very little to say about the cross-language perspective addressed in this paper, although additional material on this is available on the Council of Europe website (North and Jones 2009). It also preserves the conceptual distinction between 'standard setting' (i.e. task-based) and 'external validation', i.e. a range of validation activities focusing on the competence of learners, and based on information from a range of sources beyond the test itself. In my opinion these choices are unfortunate, as they still reinforce the impression that task-based standard setting remains the necessary and sufficient guarantee of a valid alignment process. As I have argued here, this is very far from the case.

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

Aligning Cambridge ESOL examinations to the CEFR: issues and practice

HANAN KHALIFA RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

ANGELA FFRENCH ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

The Common European Framework of Reference (CEFR) aimed to provide 'a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe' (Council of Europe 2001a:1). Today, the CEFR's influence extends beyond Europe and despite the clear statement made by its authors in relation to its use 'we have not set out to tell practitioners what to do or how to do it', (Council of Europe 2001a:xi), for many language testers, it has become imperative to make the case that their exams are aligned to the CEFR.

As mentioned earlier in the opening article to this issue, the Council of Europe has attempted to facilitate this linking exercise by providing a toolkit of resources, including a draft pilot *Manual* for relating language examinations to the CEFR. In this paper we discuss Cambridge ESOL's experience in using the pilot version of the *Manual* (Council of Europe 2003a).[1] We compare the *Manual's* approach to alignment and Cambridge ESOL's iterative and cyclic approach to alignment. The comparison will demonstrate how we can build an alignment argument that is based on existing test development and validation systems while generating evidence in line with the aims of the *Manual*. We would like to note here that at the time of writing this article, the final version of the *Manual* has been published and it has taken on board some of the recommendations we have provided in our 2008 paper, e.g. the concept of embedding procedures recommended in the *Manual* at each stage of the test development (Council of Europe 2009:9).

Approaches to alignment

A discussion on alignment cannot really take place without reference to the unique relationship the CEFR has with Cambridge ESOL Examinations in terms of their shared purposes, namely, provision of a learning ladder and proficiency framework and in terms of informing each other's evolution and development (see North 2008, Taylor and Jones 2006, www.cambridgeesol.org/what-we-do/research/cefr/empirical-perspective.html for a full discussion of this relationship).

Manual's approach to alignment

The *Manual* outlines the alignment process as a set of inter-related stages: *Familiarisation, Specification, Standardisation training/benchmarking, Standard setting* and *Validation*. Familiarisation aims at ensuring participants have detailed knowledge of the CEFR. Specification involves mapping the content and task types of an examination in relation to certain categories of the CEFR through making use of a number of illustrative scales. Standardisation includes training and benchmarking which aim to achieve a common understanding and interpretation of the meaning of the CEFR levels. Standard setting is where a decision is made with regard to cut-off scores and placing students on one of the

Table 1: Summary of linking steps recommended in the *Manual*

Familiarisation
Background reading and discussion of the CEFR levels as a whole
Sorting individual CEFR descriptors from a CEFR scale into levels
Self-assessment of own language level in a foreign language
Reconstructing CEFR tables and grids from individual descriptors
Illustrating CEFR levels with learner performance
The outcome of the familiarisation process is a narrative description reporting on the activities used and the results obtained.
Specification
Describing the examination in question: objectives, learner needs, exam papers/sections, weighting/rationale for weighting, text and task/item types, marking criteria, scoring scheme and reporting results
Content analysis of the examination profiled in relation to the categories and levels of the CEFR
The outcomes of the specification process are: • a chart profiling coverage graphically in terms of levels and categories of the CEFR • narrative description of the process • completed versions of relevant forms.
Standardisation training/benchmarking
Training with exemplar performances and test tasks to achieve a common understanding of the CEFR levels
Benchmarking local samples of performance for productive skills and local tasks/items for receptive skills
Statistical analysis of ratings
The outcomes of the standardisation process are: • examples of tasks and task-templates • samples of spoken and written performance • procedures for marking and rating • narrative explaining how CEFR standardised performance samples were exploited to benchmark local samples to the CEFR and set up moderating systems.
Standard setting
Selection of a standard setting method
Consideration of standard setting with multi-skill tests
Consideration of applying test equating methodology
The outcome of the standard setting process is a detailed record of the procedure followed and analysis applied.

[1] An earlier version of this paper was presented at the 34th Annual International Association for Educational Assessment Conference in Cambridge, 8–12 September 2008.

CEFR levels. Validation is perceived as an ongoing process to ensure the quality of each of the previously mentioned stages. To facilitate the implementation of the *Manual's* approach a set of activities, forms and illustrative examples are provided. Table 1 summarises the steps that are recommended in the *Manual* in terms of Familiarisation, Specification, Standardisation and Standard setting.

When considering the recommended linking scheme in the *Manual* and its toolkit of resources, one cannot help but admire its rigour and thoroughness. However, it seems that the *Manual* envisages the alignment process as a specific project being organised, possibly on a one-off basis, where participants are trained to carry out a set of activities, and reports are generated which constitute the evidential outcomes. This may very well be the case if no change occurs in the context within which an examination has been developed and administered or if the exam does not undergo changes. The criticality of the alignment process lies in its sustainability. The section below provides an account of how an examination board interprets the *Manual's* approach to alignment while at the same time ensures the sustainability of the process.

Cambridge ESOL's interpretation of the approach

Cambridge ESOL views the alignment of its tests to the CEFR as a key aspect of gaining an understanding of the quality of performance that is represented by a candidate's score on a particular test; as part of its communication strategy to stakeholders with regard to the meaningfulness of the results of an examination. Cambridge ESOL also views alignment as an embedded and integrated feature of its test development and validation model. The model has an ongoing iterative cycle from perceived need, through test design, trialling and administration to post exam review (for full discussion of the model see Saville 2003). The cycle allows for changes in learning, pedagogy and assessment trends, as well as in the targeted candidature, to be incorporated into an examination. Therefore, by default, the linking process is perceived as an ongoing activity rather than as a single activity at a given time. In this section, we show how we can build an argument based on the high quality of existing processes, while generating evidence in line with the aims of the *Manual*.

Familiarisation

The *Manual* perceives the familiarisation procedure as an 'indispensable starting point' before a linking exercise to the CEFR can be carried out effectively. This perception leads to the emergence of several issues, e.g., who are these participants, how many, how familiar are they with the CEFR on a familiarity-unfamiliarity continuum. Should we take their assessment experience into account or should we start with an assumed zero baseline of knowledge? What mode of delivery should the familiarisation activity take: face to face or distant?

Within Cambridge ESOL, the participants who ensure that an examination is related to the CEFR span across the organisation and beyond. There is currently a core of 60 assessment, operations and research staff who work with a network of external stakeholders on test specification, item writing, test construction, performance scale construction and on the application and use of performance testing scales. In July 2009, the network included:

- 160 personnel involved in writing materials
- approximately 25,800 oral examiners (750 of whom are team leaders, and about 95% of whom do not reside in the UK)
- 2500 writing examiners of which around 100 are team leaders.

Roles and responsibilities are at times interchangeable, e.g., a chair of one exam could be an item writer for another.

Experienced members of this large community certainly have a close understanding of Cambridge ESOL levels. For example, item writers providing materials for a particular examination are familiar with how to interpret the level, e.g., in terms of text difficulty, linguistic features, genre choice, etc. Item writer guidelines and examination handbooks provide detailed information on text selection and item writing at a certain level. Through their work with a particular proficiency level, many will have come into contact with the CEFR or at least Waystage 1990, Threshold 1990 and Vantage specifications on which the CEFR is based (van Ek and Trim 1990b, 1990a, 2001). Many will also be familiar with the ALTE Can Do descriptors which have been calibrated to the CEFR scale and which are reported in examination handbooks. For Cambridge ESOL, then, familiarisation with the CEFR is seen as a part of consolidating and building on existing knowledge. At the same time the prominence of the CEFR raises the need for a more general awareness-raising, of particular importance for newer staff and members of our stakeholder community.

One way the organisation has incorporated *Manual*-advocated familiarisation activities into its practice is through its ongoing induction and training programme; annual seminar programme, and annual team meetings. The familiarisation event may take the form of a face-to-face workshop or self-accessed materials using an electronic platform for queries and feedback. Through the familiarisation activities, Cambridge ESOL aims to foster a common understanding of the objectives and aspirations of the CEFR and its descriptive scheme and a broad awareness of the nature of the relationship between Cambridge ESOL examinations and the CEFR; to ensure a shared knowledge of differentiating features across certain level thresholds (e.g. B1/B2 and B2/C1) to enable the rating of tasks and performances across these levels.

Specification

The *Manual* has provided several forms for use in describing the examination and mapping its content onto the CEFR. These forms elicit general information on the examination, on test development, the construct being measured, marking, analysis, grading and results reporting. Cambridge ESOL describes such information in internal documents, e.g., item writer guidelines, routine test production process, standard operating procedures for exam production, and grading instructions as well as external documents such as examination handbooks or annual reports on examination performance which are available on our website. These documents specify the exam constructs and levels and explain how they are implemented very effectively and already make use of many things now incorporated into the CEFR but familiar for many years – the Waystage and Threshold specifications, for example.

Mapping of the examination content onto the CEFR does not happen as a single activity nor is it carried out by a single member of staff; it occurs at different stages of Cambridge ESOL's test development cycle and a number of internal and external personnel participate in the process bringing together a variety of expertise. It happens at the planning phase, the design phase and the development phase. At these stages specifications are produced linking candidate needs to requirements of test usefulness including frameworks of reference such as the CEFR. Decisions are made with regard to exam proficiency level, text and task parameters as well as features of performance scales which illustrate the proficiency level, marking methodology and procedures for setting grade boundaries. Pretesting and trialling take place to confirm decisions made and/or allow for modifications. The use of a socio-cognitive approach to test development has also helped Cambridge ESOL define and operationalise the measured construct in terms of the CEFR levels (see Weir 2005b).

Further explicit reference to the CEFR is being introduced into Cambridge ESOL's processes over time where this serves to complement or clarify, for example when examinations are revised and updated. Task design and scale construction for performance tests is a case in point as we see in the following section.

Table 2: Language use for Level B2

Common reference levels (Table 3, p185)	Cambridge ESOL Analytical Scales	FCE Speaking
• Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so. *(Range)*	• Uses a range of appropriate vocabulary to give and exchange views on a range of familiar topics. *(Grammar and Vocabulary)* • Contributions are relevant and there is very little repetition. *(Discourse Management)*	FCE candidates are expected to use a range of appropriate vocabulary to give and exchange views on a range of familiar topics. Hesitation may occur whilst searching for language resources (ESOL Common Scale), though this is not expected to be conspicuous. In the Individual Long Turn (Part 2), candidates are not expected to give detailed descriptions of their photographs, but rather they are expected to compare them and give their reactions to them, and students are advised to paraphrase when they do not know, or cannot remember a word (Handbook p77). They should listen carefully to the instructions which follow the words '*and say*' and read the questions above the photographs. If they do not do this they may miss the focus of the task and not produce a wide enough range of language. FCE candidates are expected to attempt some complex grammatical forms, and to use a range of cohesive devices.
• Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes. *(Accuracy)*	• Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. *(Grammar and Vocabulary)*	The assessment criteria of *Grammar and Vocabulary*, which focus in a positive manner on what candidates can, rather than cannot do, require candidates to show a good degree of control of simple grammatical forms, to attempt some complex grammatical forms, and to use a range of appropriate vocabulary.
• Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses. *(Fluency)*	• Produces extended stretches of language despite some hesitation. *(Discourse Management)*	Despite some hesitation, FCE candidates are expected to produce extended stretches of language and maintain a flow of language, with sentence and word stress generally accurately placed. The stretch of language used by the candidate should be appropriate to the task. In Part 2, each candidate is given the opportunity to engage in an individual one minute uninterrupted 'long turn'. In addition to comparing two photographs, candidates are required to focus on an aspect of the topic of the photographs, e.g. '… and say why you think the music is important to the different groups of people.' The long turn allows candidates to demonstrate their ability in organising a larger unit of discourse.
• Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc. *(Interaction)*	• Initiates and responds appropriately. *(Interactive Communication)* • Maintains and develops the interaction and negotiates towards an outcome with very little support. *(Interactive Communication)*	In the Part 3 collaborative task, the interlocutor sets up a task and then withdraws, encouraging the two candidates to take responsibility for the management of the interaction and to work together towards a negotiated outcome of the task set. Candidates are encouraged to respond to each other's contributions by agreeing, disagreeing and questioning each other, rather than just by giving information about the task (Handbook p78).
• Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution. *(Coherence)*	• Uses a range of cohesive devices. *(Discourse Management)* • Produces extended stretches of language despite some hesitation. *(Discourse Management)*	FCE candidates are expected to use a range of cohesive devices, and the individual long turn in Part 2 gives them the opportunity to show their ability to organise their thoughts and ideas, and express themselves coherently.

Exemplification from the CEFR scales and the FCE Speaking scale

The First Certificate in English (FCE) is a Cambridge ESOL examination at B2 level measuring all four skills. Performance on the Speaking paper is assessed on the following analytical criteria: Grammar and Vocabulary (G&V), Pronunciation (P), Discourse Management (DM), Interactive Communication (IC). Table 2 shows how the analytical scales in the Cambridge ESOL assessment criteria for speaking satisfy the requirements of the CEFR. Elements of Range and Accuracy (CEFR) are evidenced in the Cambridge ESOL criteria of Grammar and Vocabulary and Pronunciation. Fluency and Coherence (CEFR) are captured under the Discourse Management criterion, while Interaction (CEFR) and Interactive Communication (Cambridge ESOL) focus on very similar aspects of the performance.

The table also shows how the qualitative aspects of spoken language use as outlined in the *Manual* are satisfied through the test format, task types and assessment criteria used in the FCE Speaking test. When considering, for example, part 2 of FCE Speaking paper, we find that the task type and test focus are presented as follows in the FCE handbook for teachers (UCLES 2007a):

- **Task type and format:** An individual long turn with a brief response from the second candidate. In turn, the candidates are given a pair of photographs to talk about.
- **Test focus:** Organising a larger unit of discourse; comparing, describing, expressing opinions.

Standardisation

Standardisation involves 'achieving and implementing a common understanding of the meaning of the CEF levels' (Council of Europe 2009a:11). The *Manual* states that this involves:(a) *Training* professionals in a common interpretation of the CEFR levels using productive skills samples and receptive skills calibrated items which are already standardised to the CEF; (b) *Benchmarking* where the agreement reached at *Training* is applied to the assessment of local performance samples before moving on to *Standard setting* where the cut-off scores for the test CEF level(s) are set using a standard setting technique.

Cambridge ESOL addresses the *Manual's* Standardisation activities through its existing rigorous system of recruitment, induction, training, coordination (standardisation), monitoring and evaluation (RITCME) for its item writers and examiners (see Khalifa and Weir 2009, Shaw and Weir 2007 for a full account of the system where assessing reading and writing are concerned respectively and Ingham 2008 for a detailed discussion on how RITCME is implemented). RITCME is a continuous process and remedial actions are carried out effectively and efficiently. This system ensures that these participants have adequate professional background and receive appropriate training in the skills required. It also allows for ongoing professional development through standardisation, monitoring and evaluation of the performance of item writers and examiners of productive skills. Obligatory standardisation of writing examiners takes place prior to every marking session, and the writing samples used are evaluated by the most senior examiners for the paper. Coordination of oral examiners takes place once a year prior to the main administration session and the video samples of performances which are used are rated by Cambridge ESOL's most experienced Professional Support Leaders and Team Leaders, representing a wide range of countries and familiarity with level. The marks provided are then subject to quantitative and qualitative analysis before being approved for standardisation purposes. Materials used in the training and coordination events have already been standardised to the CEFR and some of these materials have been provided to the Council of Europe as illustrative receptive and productive samples of the CEFR levels.[2] Recently, due to exam updates, Cambridge ESOL has set up a project to select oral performance exemplifying CEFR levels (see Galaczi and Khalifa 2009 in this issue). Once again we see here how the *Manual's* recommended Standardisation activities are embedded within Cambridge ESOL practice.

Standard setting

Standard setting as described in the *Manual* is perhaps most applicable when designing a new examination or in a context when item banking methodology is not in use. Cambridge ESOL examinations are mapped onto a measurement scale common to all examinations. For receptive skills, for example, the stability of the measurement scale is achieved by an item banking methodology that is employed in the development phase where new items are pretested and calibrated using anchor items to monitor exam difficulty. The calibrated items are then stored in the Cambridge ESOL Local Item Banking System (LIBS) where each item has a known difficulty and accordingly examination papers are constructed to a target difficulty on the CEFR A2–C2 continuum and can be graded accordingly to a high degree of precision. This is better described as *standard-maintaining* rather than *standard setting* given that the standard is a stable one which is carried forward (see North and Jones 2009 for a discussion on this). The current rationale for the standard of the objective papers owes something to an essentially normative view of skill profiles in a European context (as, probably, does the CEFR), and something to the progression depicted by the common measurement scale, which can be represented as a rational 'ladder' of learning objectives.

Validation

Certain aspects of *Validation* as described by the *Manual* in 2009 relates to a very wide range of activities in the different phases of Cambridge ESOL's test development and administration cycle. It is related to:

- the planning stage where data is gathered on the targeted candidature through reliable and valid data collection instruments
- the design and development stages where data is collected to ensure that the examination is reliable, valid, and practical

[2] See www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html

- the test administration stage where quality assurance procedures are in place to ensure fair administration procedures
- the post exam and review stage where data is collected on the washback effect of the examination.

Cambridge ESOL follows a socio-cognitive approach towards test development and validation. This approach considers and provides evidence for the following components of validity: cognitive, contextual, scoring, criterion-related and consequential. Linking to the CEFR is embedded in all of these. Questions like 'what makes a B2 reader differ from a B1 reader in terms of cognitive ability?' or 'what type of task is most suited to assess the ability to integrate information form a variety of reading texts which is a mark of a C1/C2 reader? are posed, answered and documented (See for example Shaw and Weir 2007 on examining writing; Khalifa and Weir 2009 on examining reading).

Within Cambridge ESOL, validation is carried out through a variety of activities:

- Statistical analyses of objective items before (pretesting) and after live sessions. This includes the use of anchor tests, and information about candidates gathered each session via candidate information sheets.
- Qualitative analysis of Writing and Speaking tasks before (trialling) and after live sessions which is documented in examiner, professional support leader and annual validation reports.
- Statistical analysis of Writing examiners' marking tendencies and monitoring via the professional support leader system through the entire marking period, in addition to a systematic 'marks collection' exercise in Speaking coordination and in the monitoring process in live sessions.
- A research programme investigating issues like exam comparability, version equivalence and so on.

Conclusion

The *Manual* aims to assist examination providers to 'develop, apply and report transparent, practical procedures in a cumulative process of continuing improvement in order to situate their examination(s) in relation to the CEFR' (Council of Europe 2009:1). The *Manual* appears to envisage the provision of different types of evidence demonstrating examination alignment to the CEFR and showing the quality of the methodology. Nevertheless, it is encouraging to see that the authors of the *Manual* do not see it as the 'sole guide to linking a test to the CEFR and [that] there is no compulsion on any institution to undertake such linking' (ibid.:1). They appear to envisage that users will apply *Manual* procedures rationally and selectively and through a reflection on this application, users contribute to a body of knowledge and experience and add to the suggested techniques. However, in order to sustain alignment claims it is imperative to bring explicit CEFR reference into test providers' practices, processes and documentation on an ongoing basis coordinated for practical purposes with revisions and updates. It is worth noting here that the CEFR as a concept appeared to function quite well in the past without extensive underpinning from measurement theory and statistics. However, these are becoming more and more important as attempts are being made to validate aspects of the CEFR empirically (North and Schneider 1998) and to align examinations to it (Kaftandjieva 2004).

As pressure to use the CEFR in a regulatory way increases, we need to caution here that while frameworks carry certain benefits to a variety of stakeholders, e.g., facilitating selection from a range of examinations, they also have limitations. As Taylor (2004:5) states 'they risk masking significant differentiating features, they tend to encourage oversimplification and misinterpretation, and there is always a danger that they are adopted as prescriptive rather than informative tools'. The purpose of any linking exercise is to provide a framework of how tests and levels relate to each other in broad terms within a common frame of reference. This is of particular value to end users especially in the globalised world in which we now live. The major challenge, therefore, for language testers, at least in Europe, is to begin to look explicitly at direct cross-language comparison. This will need new methodologies and kinds of evidence, but provides the best hope of a better answer to the question: 'How does my B1 *compare* with your B1?, In what way do they vary?'

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

19 October 2009: 4th Cambridge Assessment Conference

The 4th Cambridge Assessment Conference will take place in October, on the theme of *Issues of control and innovation: the role of the state in assessment systems*. Hosted by the Cambridge Assessment Network at Robinson College, Cambridge, UK, the conference will bring together more than 200 public policy experts, educationalists and assessment specialists. Keynote speakers include: Professor Alison Wolf (King's College London) and Professor Robin Alexander (University of Cambridge).

For further details or to book a place, please visit www.assessnet.org.uk/annualconference

An exploratory study into linking young learners' examinations to the CEFR

SZILVIA PAPP RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL
ANGELIKI SALAMOURA RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

Cambridge ESOL's Young Learners English (YLE) tests were originally designed in 1997 for children aged 7–12 to build a bridge from beginner to *Waystage* level (van Ek and Trim 1998b). The YLE tests are offered at three levels: Starters, Movers and Flyers. Flyers, as the highest of the three, was designed to be at approximately the same level as Cambridge KET, i.e. CEFR level A2, Movers at a level below, i.e. CEFR level A1, while Starters at near-beginner level, i.e. below CEFR level A1. The CEFR was never intended to be a framework of reference for young children's language learning, teaching and assessment. Nevertheless, recent validation work has explored the possibility of formally aligning YLE tests to the CEFR based on quantitative and qualitative research. This article reports on the qualitative strand.

Methodology

The Research and Validation Group of Cambridge ESOL commissioned an external expert to work individually and with YLE Chairs, Subject Manager, Subject Officers, Assistant Subject Officers, and Validation Officers (henceforward referred to as the YLE team) on aligning the YLE tests at Flyers, Movers and Starters levels to the CEFR as laid out in the procedures in Chapter 4 of the *Preliminary Pilot Version of the Manual* for relating language examinations to the CEFR (Council of Europe 2003a). The external expert's and YLE team's task was to:

- Map the construct as defined in YLE test specifications (i.e. the YLE Handbook and Item Writer Guidelines) and YLE task specifications to Chapters 4 and 5 of the CEFR (Council of Europe 2001a).
- Rate sample spoken performances and sample tasks taken from each level of the YLE tests using relevant CEFR scales.
- Identify salient features of language use and competence by YLE candidates at all three YLE levels by making comparisons with the relevant sections of Chapters 4 and 5 of the CEFR (ibid.).

The purpose of the workshop organised for the linking exercise was two-fold:

1. To familiarise the expert and the YLE team with the CEFR scales, in particular bands of descriptors relevant for the YLE tests, and to train them in rating sample Council of Europe performances in CEFR levels, especially those that are relevant for the YLE tests (pre-A1–B1).
2. Using the training received, to rate sample YLE oral performances (benchmarking) and sample YLE tasks (standard setting) using relevant CEFR scales and hence contribute to the verification of the alignment of YLE tests to the CEFR.

Research questions

The following research questions were pursued.

1. What is the demonstrable link between YLE test specifications, task specifications, YLE candidate performances and the relevant CEFR scales and descriptors in terms of proficiency levels and activities, strategies and competences YLE candidates can be expected to do or have?
2. How do the YLE tests embody and reflect the CEFR in terms of proficiency levels and what activities, strategies and competences can YLE candidates be expected to do or have?
3. What are the salient features of YLE candidate performance that reflect the CEFR in terms of proficiency levels and activities, strategies and competences YLE candidates can be expected to do or have?
4. Which additional salient features can be identified in YLE candidate performances that are not currently covered in the CEFR?

As can be seen in these research questions, the main aim of the study was to see whether the CEFR can be inspirational in a YLE context (cf. Jones and Saville 2009).

Procedures and materials

In this section, we detail the procedures and materials used in the Familiarisation, Specification and Standardisation stages of the linking exercise as suggested by the *Preliminary Pilot version of the Manual* for relating language examinations to the CEFR (Council of Europe 2003a; henceforward *Pilot Manual*) available to the team at the time.

Familiarisation

Familiarisation activities were carried out both prior to and at the face-to-face workshop. Prior to the workshop, the external expert and all members of the YLE team were asked to fill in 3 CEFR worksheets that have been developed by Cambridge ESOL and adapted for the purposes of this linking exercise. Worksheets 1 and 2 provided a thorough theoretical introduction to the CEFR and its relationship with Cambridge ESOL exams. Worksheet 3 included two hands-on activities. The first one was a self-assessment task in which the YLE team was asked to self-assess their ability in a foreign language using the CEFR Self-assessment Grid in Table 2 of the CEFR (Council of Europe 2001a:26–7) and mark key words in the descriptors of their chosen level and its adjacent levels in order to better understand the salient features in these levels (based on Familiarisation activity *c* of the *Pilot Manual,* Council of Europe 2003a:26). The second activity was a descriptor-sorting task in which participants were asked to classify mixed-up CEFR descriptors to the appropriate level and justify their choice based on the CEFR Global Scale, i.e. Table 1 of the CEFR (Council of Europe 2001a:24) (based on Familiarisation activity *d* of the *Pilot Manual*, Council of Europe 2003a:27). This exercise focused on descriptors from the CEFR levels A1–B1 only, i.e. those levels closest to the YLE proficiency levels, and used the following tables from the *Pilot Manual* (Council of Europe 2003a):

- Speaking: Global Oral Assessment Scale (Table 5.4, p78)
- Writing: Written Assessment Criteria Grid (Table 5.8, p82)
- Reading and Listening: Relevant Qualitative Factors for Reception (Table 4.3, p53).

For all activities, the YLE team was asked to send their responses back to the project coordinator for collation, and they were subsequently provided with keys to check their answers. For the descriptor-sorting activity, they were also asked to keep notes about their descriptor classification and to prepare to discuss them at the workshop.

In the first 1-hour session of the face-to-face workshop, the responses to the descriptor sorting exercise done previously at home were discussed and agreed upon. Particular emphasis was put on those descriptors which caused classification problems to more than one rater.

Specification

This stage involved filling in forms A1–A23 of the *Pilot Manual* in order to map the content of YLE tests in relation to the CEFR levels and provide evidence of the internal and external validity of YLE. The forms were completed making use of information available in the YLE Handbook and internal Item Writer Guidelines as well as YLE task specifications, mapping features of language use and competence covered in the YLE tests to Chapters 4 and 5 of the CEFR. Forms A2 and A8–A23 were then filled in for each level of YLE. These forms lay out evidence for external validity.

Standardisation

This part comprised a number of activities, summarised below.

Training in rating oral performances using CEFR scales

The expert and members of the YLE team (altogether 6 raters on the day of the workshop) received training in rating oral performances across the A1 to B1 levels, i.e. the CEFR levels closest to the YLE proficiency levels. Table 1 lists the three short oral performances that were used for this purpose. The performances were selected from the Eurocentres DVD which contains CEFR calibrated samples (Council of Europe 2003c). At the time of the linking exercise, this was the only source that included an A1 performance. Also, the samples' duration was close to that of the oral performances elicited in the YLE tests. Workshop participants had to watch the performances and rate them using the Global Oral Assessment Scale of the *Pilot Manual* (Council of Europe 2003a:78), which they had also worked with in the Familiarisation descriptor-sorting activity. In addition, they were also given the choice of using an analytic scale, the Oral Assessment Criteria Grid of the *Pilot Manual* (ibid.:79), if they thought it would facilitate the accuracy of their ratings.

In these sample performances, the production task is to talk about a topic in a sustained, coherent monologue that is semi-prepared. The task shows what the learner in question can do given an opportunity to reflect on what they want to say.

Table 1: Materials used in the training rating activity (Speaking)

A. Speaking samples *(from Council of Europe 2003c Eurocentres DVD)*

No	Candidate	CEFR level	Task	Discourse type	Duration
1.	Marcel	A2	"My home"	Monologue	~ 4min
2.	Micheline	A1	"Last weekend"	Monologue	~ 4min
3.	Renate	B1	"My flat"	Monologue	~ 4min

B. CEFR Tables/Scales

Table 5.4: Global Oral Assessment Scale and Table 5.5: Oral Assessment Criteria Grid of the Manual (Council of Europe 2003a:78)

After watching each performance, participants consulted the CEFR scale(s) and filled in a rating form based on form B3 of the *Pilot Manual* (Council of Europe 2003a:80) with their ratings and justification/rationale using key words and notions from the CEFR scales. At the end of the rating exercise a general discussion followed where participants were presented with the correct ratings from the DVD and were asked to explain – according to their understanding – what the key differences between levels were and how they arrived at their ratings. They were also given the reference pages from the DVD's accompanying documentation for the performances they worked with. To aid understanding of the performances' link to the framework, these pages explicate the features of each performance in relation to the CEFR levels.

Benchmarking – rating of YLE oral performances

The expert and members of YLE team now trained in the use of relevant CEFR oral scales were asked to rate one typical YLE oral performance at each level, using the same scales as in the training session. The YLE samples were chosen to be typical, solid sample performances at each level, ensuring their representativeness of the YLE candidature in terms of ability, age and L1. It was decided to use average performances, not performances that received full marks in order to avoid a ceiling effect in their rating in the workshop. It was also agreed that candidates' full exam performances were to be rated rather than partial performances, even though the YLE speaking tests are not comparable in length at each level of the YLE tests.

The following performances from the 2007/2008 YLE Standardisation DVDs were identified for benchmarking and were presented in the following random order for rating:

1. Xavi – Movers
Tasks: 1. Find the differences: Stairs
2. Picture story: A kangaroo helps Tony

2. Maria – Starters
Tasks: 1. Object cards
2. Scene picture: Sea

3. Simon – Flyers
Tasks: 1. Find the differences: River
2. Information exchange: Rooms
3. Picture story: Mum's great ideas
4. Personal questions: Your school

The rating procedure was exactly the same as that in the training session, followed by a group discussion about the features of each performance and how they relate to the CEFR levels. Again, participants were given the ratings and rationale from the Guidelines for Oral Examiner Training and Coordination for each YLE level.

Standard setting – rating of YLE Reading/Writing and Listening tasks

The Reading/Writing and Listening tasks for the standard setting activity were selected to be equivalent in terms of difficulty. Once identified, the tasks were presented in random order for rating in the workshop.

A variant of the Angoff method of standard setting was adopted. The YLE team was asked to answer the following question: 'At what CEFR level can a test taker already answer the following task correctly?'. The panellists had to reflect on the minimum competence required for successful completion of each task in terms of the CEFR levels. Participants were asked to think at task rather than item level, as the *Pilot Manual* suggests, because the majority of the YLE tasks do not contain discrete items. Workshop participants recorded their responses and any other comments if they wished in a Task Rating Form based on Form B5 of the *Pilot Manual* (Council of Europe 2003a:86).

Regarding the scales used in the standard setting session, the YLE team was presented with the Relevant Qualitative Factors for Reception scale, i.e. Table 4.3 of the *Pilot Manual* (Council of Europe 2003a:53), as suggested by the *Pilot Manual*. However, most participants found it difficult to use this scale for standard setting purposes as they felt that the descriptors were not relevant for capturing the features of the YLE tasks. Instead the panellists agreed to use the Overall Reading Comprehension Scale of the CEFR (Council of Europe 2001a:68) and the Overall Listening Comprehension Scale of the CEFR (ibid.:66) which were deemed to be more appropriate for the purposes of the session. The YLE Reading/Writing tasks require minimal writing skills (mainly copying) and therefore no scale was appropriate for assessing YLE writing skills, which all participants felt were therefore, by definition, at pre-A1 level.

Table 2: Raters' scores in the descriptor-sorting activity: Responses matching the target CEFR level

Scales	Descriptors	Mean	Min	Max	SD	Percentage
Speaking	8	6.83	4	8	1.60	85%
Writing	9	7.33	5	9	1.63	82%
Reading and Listening	15	11.83	11	14	1.17	79%
Overall	32	26	20	31	4	81%

Results

Familiarisation – CEFR descriptor-sorting activity

Table 2 presents the average raw scores and percentages of participants' responses which placed CEFR descriptors in the correct CEFR level in the pre-workshop descriptor-sorting activity. The data were scored here in a dichotomous way – 1 for correct placement, 0 for wrong placement. Exact agreement in all skills is satisfactory as it has an average of at least 79%. Not surprisingly, agreement was slightly better in the productive than in the receptive skills during the familiarisation activity.

Average values of inter-rater reliability during the familiarisation exercise are shown in Table 3. The correlations reported are Spearman rank order coefficients and they were estimated for all possible pairs of raters. With the exception of some minimum correlations in Speaking, all other Spearman correlations in this table are significant at $p \le 0.05$. These inter-rater reliability values together with the high alpha values indicate a satisfactory agreement among raters on ranking the descriptors from the lowest (A1) to the highest level (B1) during the familiarisation activity.

Table 3: Inter-rater reliability and internal consistency: Summary statistics

Scales	Inter-rater reliability			Alpha
	Mean	Min	Max	
Speaking[1]	0.74	0.49	1	0.95
Writing[2]	0.80	0.62	1	0.96
Reading and Listening[2]	0.77	0.53	0.94	0.95
Overall[3]	0.75	0.59	0.93	0.95

1 Only 7 correlation coefficients (out of a total 15) are significant at $p < .05$.
2 All correlation coefficients are significant at $p < .05$.
3 All correlation coefficients are significant at $p < .01$.

The lower inter-rater reliability values in Speaking (0.49) were due to two raters. Without these two raters the average correlation in Speaking is 0.95 and the minimum 0.9.[1] Despite this, it was decided that ratings from these two judges would not be excluded from the analyses of the

[1] The fact that participants had to rate individual descriptors that came from the same level rather than descriptions of whole levels may have contributed to this rather low inter-rater agreement.

Training, Benchmarking and Standard Setting data as their ratings were not found to be significantly different from the ratings of the rest of the judges in those later exercises. This also indicates that the discussion which followed the descriptor-sorting (familiarisation) activity was efficient in gaining a common understanding of the CEFR scales and descriptors.

Standardisation

Training in rating oral performances using CEFR scales

The ratings given by the panellists during the Training, Benchmarking and Standard Setting sessions contained labels such as pre-A1, the + mark appended to levels, and portmanteau ratings for borderline cases, such as A1/A2. These labels needed to be converted into numbers, as shown in Table 4, in order to be statistically analysed. The resulting numerical scale is an arbitrary scale for measurement purposes only and does not imply equal distance between the different levels and sublevels awarded by the raters in this study[2].

Table 4: Conversion of labels representing CEFR levels given by panellists into a numerical scale

CEFR level	Scale
pre-A1	**1**
pre-A1+	1.25
pre-A1/A1	1.5
A1-	1.75
A1	**2**
A1+	2.25
A1/A2	2.5
A2-	2.75
A2	**3**
A2+	3.25
A2/B1	3.5
B1-	3.75
B1	**4**

Due to the small number of CEFR levels represented by the performances and tasks in this study (three learner performances or set of tasks – one for each of the YLE levels) in the Training, Benchmarking and Standard Setting sessions, a standard reliability index (e.g. Spearman *rho* correlation, or alpha) could not be calculated for raters. Following Bachman, Davidson and Milanovic (1996) as well as Clapham (1996:151), we have therefore used the Weighted Rater Agreement Proportion (WRAP) statistic which is suitable for measuring the proportion of rater agreement on each item when there are few items to be rated. In our context, if all six raters agree, WRAP is 1 (6/6); if five raters agree, WRAP is 0.83 (5/6), etc. If, however, one rater is half a mark away from the majority of the raters, then this rater contributes 0.5 mark to the total; for instance, if five raters agree and the sixth is half a mark away, WRAP is 0.92 (5.5/6). Similarly, if a rater is a quarter of a mark away from the majority, then this rater contributes 0.75 marks to the total.

Table 5: Training: Summary statistics and rater agreement from the speaking rating activity

Learner	Mean	Median	Mode	Min	Max	SD	WRAP[1]
Marcel (A2)	3.17	3	3	3	3.5	0.26	0.83
Micheline (A1)	1.96	2	2	1.75	2	0.10	0.96
Renate (B1)	4	4	4	4	4	0	1

[1] WRAP = Weighted Rater Agreement Proportion (Bachman, Davidson and Milanovic 1996, Clapham 1996:151)

Table 5 shows very high agreement among raters regarding the target CEFR levels of the three performances in the Training session (Marcel A2, Micheline A1, Renate B1). This shows a very good understanding of the features of oral performances across the A1–B1 levels by the raters.

Benchmarking – rating of YLE performances

In the Benchmarking session, as Table 6 shows, there was also high agreement about the CEFR level of Maria (pre-A1) and Simon (A2). The agreement for Xavi was lower (71%). On average, he was rated as a weak A1 (A1-) but the most frequent rating (the Mode) was A1. Although Xavi was rated as a weak rather than a clear A1 performance, what is important is that the raters saw a clear progression between CEFR levels from Starters (pre-A1) through to Movers (A1) to Flyers (A2) as exemplified by the three sample performances from YLE candidates.

Table 6: Benchmarking: Summary statistics and rater agreement from the speaking rating activity

Learner	Mean	Median	Mode	Min	Max	SD	WRAP
Xavi (YLE Movers)	1.71	1.88	2	1	2	0.40	0.71
Maria (YLE Starters)	1.13	1	1	1	1.75	0.31	0.96
Simon (YLE Flyers)	2.88	3	3	2.5	3	0.21	0.83

Standard setting – rating of YLE Reading/Writing and Listening tasks

As can be seen in Table 7, rater agreement was high with regard to the Movers and Flyers Listening tasks: 75% for the Movers Listening tasks (A2) and 83% for the Flyers

Table 7: Standard setting: Summary statistics and rater agreement from rating sample YLE Listening tasks

Learner	Mean	Median	Mode	Min	Max	SD	WRAP
Task 1 (YLE Starters)	2.33	2.25	2	2	3	0.41	0.67
Task 2 (YLE Movers)	2.92	3	3	2	3.25	0.47	0.75
Task 3 (YLE Flyers)	3.25	3.25	3.25	3	3.5	0.22	0.83

[2] In the case of labels Pre-A1/A1, A1/A2, A2/B1, panellists thought these were borderline cases. It is worth noting that the labels panellists chose to assign contain subdivisions according to their perceptions of the level of tasks and performances. These labels are possibly more meaningful within a teaching and learning context.

Listening tasks (A2+, i.e. a strong A2). Agreement was lower for the Starters Listening tasks (67%). On average, the Starters Listening tasks were rated to be strong A1 (A1+) although the most frequent rating was A2. Again, there is a clear progression of CEFR levels among the YLE Listening tasks, from A1+ to A2+, although the range is only 1 CEFR level rather than 3 levels as was the case with the oral performances.

The ratings in Table 8 refer to the Reading component of the YLE Reading and Writing tasks only. The writing skills required by the YLE tasks are mainly spelling and copying skills, or producing short phrases at the most, rather than producing any extended written text. The general consensus was that as a result, the Writing skills that are required and elicited in YLE tests can only be judged as being at pre-A1 level across all YLE levels if we measure them with the help of the current Writing scales in the CEFR, which were not designed for young children. In the Reading tasks rater agreement is lower than for the Listening tasks and the oral Speaking performances (58% for Starters, 75% for Movers and 63% for Flyers). This may be attributed to the fact that the raters found it difficult to map the descriptions of the CEFR overall reading scale to the YLE tasks. With regard to both Listening and Reading tasks, Milanovic (2009) in his opening article in this issue referred to the difficulty of using standard setting methodology with the current illustrative scales in the CEFR, especially for the receptive skills with them being latent constructs. Our example is a case in point. On average, the Starters Reading tasks were rated as A1/A2, and both the Movers and Flyers Reading tasks as A2+, i.e. there does seem to be progression from Starters to Flyers across the CEFR scale in terms of reading comprehension tasks. However, since agreement was rather low, these ratings should be interpreted with caution.

Table 8: Standard setting: Summary statistics and rater agreement from rating sample YLE Reading and Writing tasks (ratings for the Reading component only)

Learner	Mean	Median	Mode	Min	Max	SD	WRAP
Task 1 (YLE Starters)	2.58	2.75	3	2	3	0.49	0.58
Task 2 (YLE Movers)	3.25	3.25	3	3	3.5	0.22	0.75
Task 3 (YLE Flyers)	3.21	3.38	3.5	2	4	0.68	0.63

Verifying YLE alignment to the CEFR

This section contains reflections on the use of the CEFR to describe YLE performances and tasks. Table 9 shows the mean CEFR ratings given by the expert in the specification stage and by the YLE panel in the standardisation stage across all three levels in the YLE exam. The expert gave an overall rating to Starters as below A1 (Listening A1, Reading and Writing pre-A1, Speaking pre-A1), to Movers as A1 and to Flyers as A2.

The analyses from the expert judgement and the linking workshop show a clear progression of the YLE exams across the CEFR levels. This progression spans almost three CEFR levels, i.e. three numerical units in Table 9, for oral production (Speaking) and about one or one and a half CEFR level for perception (Listening and Reading). Overall, the average numerical figures for the successive levels of the YLE exam show clear differentiation of proficiency levels in CEFR terms.

Table 9: CEFR ratings by the expert and the YLE workshop panel across all YLE levels

Skill	YLE Starters	YLE Movers	YLE Flyers
Speaking	1.06	1.85	2.94
Listening	2.17	2.46	3.13
Reading	1.79	2.63	3.10
Average (all skills)	1.67	2.31	3.06
CEFR level[1]	Pre-A1/A1	A1+	A2

[1] For the conversion of CEFR ratings into a numerical scale and vice versa, see Table 4.

Both the expert and the YLE panel rated the Writing component in the YLE tests as pre-A1 if we measure them by the current version of the CEFR. Again, this is a direct consequence of the fact that the writing skills required in YLE are primarily spelling and copying and writing words or short phrases rather than producing coherent written discourse. Therefore Table 9 does not include the Writing ratings in the calculations for the average CEFR level of each YLE test as the writing skills described in the CEFR scales as they currently stand do not fully reflect the type of writing skills which are suitable for the cognitive and literacy development of young learners tested by the YLE exam.

Finally, Table 9 shows that the expected comprehension levels are higher than the expected production levels across all YLE exams in terms of the CEFR. This observation was also born out and corroborated in the discussion that followed the rating of the YLE oral performances. What emerged as a consensus is that the productive tasks assume evidence of knowledge and skills at a lower level in CEFR terms than the receptive tasks in YLE. Therefore Listening tasks require a higher level of performance than Speaking tasks, similar to Reading tasks being aimed at a higher level than Writing at each level of YLE. This in turn points to a jagged profile for children at each level in current CEFR terms. However, it needs to be noted that this jagged profile is even more salient if we measure children's language abilities in terms of the current version of the CEFR, which again, was not designed with young learners in mind. Nevertheless, this finding is not surprising, especially in the case of writing, as the YLE tests were developed with the assumption that there is a literacy lag in children until at least the age of 9 (Cameron 2001). The YLE measurement scales were designed to overlap (see Jones 2002), which shows an acknowledgement of uneven development of skills in children. The measurement scales offer a way of dealing with such observed 'jagged' profiles.

Describing YLE exams using the CEFR

The analysis of feedback from the expert and the YLE team show that the panel felt that the CEFR has its limitations, especially when it is applied in contexts for which it was not intended, such as children' language learning and use. These limitations were identified, discussed and recorded during the linking exercise and afterwards in elicited feedback on the activities before, during and after the workshop. These points are summarised below.

Firstly, the panellists pointed out that the CEFR descriptors are not fine-tuned enough to rate tasks at word or even phrase level, something that is characteristic of very low level exams. For instance, it was clear that some YLE tasks, especially Writing but also to some extent Speaking tasks, do not elicit enough productive language to be rated at CEFR levels. There are some open-ended tasks in Speaking, but it depends on children's personality and familiarity with the task and with the interlocutor whether children take opportunities that are offered to them for extended talk. Similarly, the most open-ended task that requires children to write creatively is writing a postcard, where children are asked to fill in gaps in the text with their own words and phrases.

Rating depends on a clear understanding of the CEFR descriptors and reaching consensus on what a certain qualifier such as 'simple' entails. The panellists felt that, in some cases, the descriptors were ambiguous. For instance, panellists that the CEFR grading between 'understanding', and 'understanding with reasonable accuracy' should actually be changed.

Participants felt that even though the CEFR's ethos is to describe what activities, strategies and competences learners can do or have, at the lower levels of pre-A1, A1 and A2, some of the descriptors are still phrased in negative terms, such as when referring to performance constraints or errors and mistakes (e.g. '*Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord*', A2 descriptor in Overall spoken interaction, CEFR, Council of Europe 2001a:74). This contributes to the difficulty of linking CEFR descriptors to the YLE exams, the specifications of which are phrased in positive terms.

Some descriptors were felt to be more relevant for YLE contexts, such as those relating to 'routines', since routines are what young learners predominantly use in the classroom. However, some other descriptors were found to be less relevant for YLE contexts, especially those that relate to domain or topic/theme of language use, such as 'job', 'studying' or 'current affairs' are clearly not relevant for young learners. Moreover, it was felt that most of the wording in the speaking scales had been adapted to construct the writing scales, which have less relevance to young learners' competences. Since the YLE tests were designed with the assumption that children's literacy skills are usually less developed than their oral skills, the reading and writing scales of the CEFR were not felt to be appropriate for young learners.

In relation to pragmatic skills, young learners typically learn and use English within the classroom context. Pragmatic knowledge such as awareness of politeness conventions come through teaching and the materials learners are exposed to, not naturalistic exposure to the target language.

The tasks designed for the YLE tests assume a different cognitive development than the one assumed in the CEFR, even at A1 and A2 levels. For instance, receptive skills are developed earlier and are stronger in children than productive skills, a developmental pattern characteristic of both L1 and L2 development.

Some frequent terms at the lower A1 and A2 CEFR levels, such as 'familiar matters' and 'familiar'/'high frequency words', need to be interpreted differently for young learners in comparison to young adult and adult learners for whom the CEFR was originally designed. For children, familiar matters include the world of imagination and fantasy. Moreover, 'high frequency words' within the YLE context means those words that are presented in the YLE vocabulary list[3], and they do not necessarily bear any connection to the real frequency of use of lexical items in English. This is especially relevant for some of the themes/topic areas, such as animals and stories.

The team concluded that some CEFR descriptors in the scales for Speaking and Listening were appropriate in general and could be used with adaptations for the YLE context. The scales for Reading and Writing would need further work to be adapted to reflect children's language learning and use. This is especially the case for current CEFR descriptors for Writing, which were found to be mostly inappropriate. For writing, totally new descriptors, mainly to do with spelling and handwriting, would need to be developed to reflect children's emerging writing skills.

Some methodological and practical issues

Participants' feedback also raised a couple of methodological points regarding the tasks and materials used in this linking exercise. First, it was felt that when more than one scale is used, it would be desirable to collate descriptors in one table for each of the skills rather than having to consult several tables. Second, the Relevant Qualitative Factors for Reception scale, i.e. Table 4.3 of the *Pilot Manual* (Council of Europe 2003a:53), which the *Pilot Manual* recommends for use with standard setting, was not deemed appropriate for the YLE standard setting exercise. It was felt that the descriptors did not adequately cover the reading and listening skills and functions required in the YLE exams. The Overall Listening Comprehension and Reading Scales of the CEFR (Council of Europe 2001a:66,68) were found to work better in this context. Third, although participants worked with individual descriptors from all the CEFR scales before the workshop (in the descriptor-sorting activity), some felt that seeing and studying the whole scales prior to the workshop would ease the amount of materials they had to deal with during the workshop. Otherwise, the YLE team found all pre-workshop and workshop activities useful and beneficial. In particular, they found the focus on the three CEFR levels (A1, A2, B1) that are closest to the YLE levels and the juxtaposition of their features in all activities very effective.

[3] Teachers who prepare learners for the YLE tests would most likely consult the vocabulary list published in the YLE Handbook when preparing learners for the YLE exams, see www.CambridgeESOL.org/resources/teacher

Salient features of YLE performances and tasks as reflected in the CEFR

From all these observations and comments it is clear that participants found it a difficult task to map YLE performances and tasks against the CEFR scales and descriptors. The next step was to find out what exactly is shared between YLE performances and tasks and the CEFR and what would need to be added to the CEFR to reflect the characteristic features and contexts of young learners' language learning and use.

First, all tables at the relevant levels (A1 and A2) in the CEFR and the *Pilot Manual* were reviewed to see how they reflect YLE performances and tasks. This work yielded the following results. The CEFR defines salient features as 'entries at each level [that] describe selectively what is seen as salient or new at that level' (2001:37).

Salient features that were found to differentiate Movers (A1) from Starters (pre-A1) are:

A1 (Movers) is the point at which children can:

- Interact in a simple way, initiate and respond to simple statements in areas of immediate need or on very familiar topics (ask and answer simple questions about themselves, where they live, people they know, and things they have, etc.).

  rather than

- Rely purely on a very finite, rehearsed repertoire of phrases, frequently-used routines and patterns limited to the performance of isolated tasks in specific situations, i.e. a list of pedagogic tasks in a primary school setting.

Therefore, these latter features are salient features of pre-A1 (Starters) candidates' performance.

Salient features that were found to distinguish Flyers (A2) from Movers (A2) are:

A2 (Flyers) reflects *Waystage* Level, where children can:

- Handle social functions (greet people, ask how they are and react to news; handle very short social exchanges; ask and answer questions about what they do at school and in their free time; make and respond to invitations; discuss what to do, where to go and make arrangements to meet; make and accept offers).
- Perform simple classroom tasks in English.

A2+ (Flyers) reflects strong *Waystage* performance, plus:

- More active participation in conversation.
- An ability to sustain monologues.

It needs to be noted that the salient features above relate only to speaking and listening interaction and do not describe reading or writing skills. For a comprehensive list of salient features of children's competencies, salient features of their reading and writing abilities, activities and strategies should be covered as well. Writing, as pointed out above, is not tested in YLE in any significant way. This is justified in that the target age group may not yet be very confident writers in their own L1, and in terms of language learning it is better to focus on listening and speaking as well as reading in a way that reflects more the way children learn their first language. However, salient features of children's emerging writing abilities (copying, handwriting, spelling and other enabling skills) could also be described for this age group.

It was found that children at Flyers (A2, *Waystage*) level *cannot* handle other salient features that the CEFR lists, for instance: getting out and about (make simple transactions in shops, post offices or banks; get simple information about travel; ask for everyday goods and services). This is because these contexts do not reflect the realities of young learners' daily lives.

Also, the salient features that distinguish B1 performance from A2 (Flyers) are clearly *not* relevant for child learners:

B1 reflects *Threshold* Level, where learners can:

- Maintain interaction and get across what they want to (give or seek personal views and opinions in an informal discussion with friends; express the main point they want to make comprehensibly; keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production).
- Cope flexibly with problems in everyday life (deal with most situations likely to arise when making travel arrangements through an agent or when actually travelling; enter unprepared into conversations on familiar topics; make a complaint).

B1+ reflects strong *Threshold* Level, plus the exchange of quantities of information

These salient features at B1 level are not things that YLE candidates could handle for linguistic and age-related reasons.

Conclusion

In this section we address the research questions set out at the beginning of the linking exercise:

1. What is the demonstrable link between YLE test specifications, task specifications, YLE candidate performances and the CEFR scales and descriptors in terms of proficiency levels and activities, strategies and competences YLE candidates can be expected to do or have?

As a result of this study, it has been demonstrated that the original link between YLE and the *Waystage* learning objective (van Ek and Trim 1998b) and the levels below that, i.e. the unpublished *Breakthrough* level (Trim 2001b) can indeed be verified. That is, YLE tests have been verified as being at pre-A1 (Starters), A1 (Movers) and A2 (Flyers) levels with the exception of the Writing element in the Reading and Writing papers which was classified as being pre-A1 level according to the current version of the CEFR across all three YLE levels. However, the activities, strategies and competences that YLE candidates can be expected to do or have are clearly different from those of adult learners, as laid out in the description, specification and standardisation forms resulting from this linking exercise.

2. How do the YLE tests embody and reflect the CEFR in terms of proficiency levels and what activities, strategies and competences can YLE candidates be expected to do or have?

The CEFR was not intended to be a framework of reference for young children's language learning, teaching and assessment. It clearly has its limitations when used to account for aspects of children's language learning needs and contexts. Therefore, any linking exercise of children's tests to the CEFR can only be indirect and claims of such linking interpreted only with full awareness and appreciation of the evidence marshalled in favour of the claimed link. Having said that, as we pointed out at the beginning, Cambridge ESOL's YLE tests are mapped to the CEFR levels pre-A1, A1 and A2 by their original design prior to launch in 1997 and in terms of evidence collected to date, including the results of this qualitative linking exercise as well as the summary of the quantitative linking exercises carried out since 2000 (Papp 2008).

3. What are the salient features of YLE candidate performance that reflect the CEFR in terms of proficiency levels and activities, strategies and competences YLE candidates can be expected to do or have?

Salient features of **Starters (pre-A1)** candidates' performance:

- Ability to rely purely on a very finite, rehearsed repertoire of phrases, frequently used routines and patterns limited to the performance of isolated tasks in specific situations, i.e. a list of pedagogic tasks in a primary school setting.

Salient features of **Movers (A1)** candidates' performance:

- Ability to interact in a simple way, initiate and respond to simple statements in areas of immediate need or on very familiar topics (ask and answer simple questions about themselves, where they live, people they know, and things they have, etc.).

Salient features of **Flyers (A2)** candidates' performance:

- Ability to handle social functions (greet people, ask how they are and react to news; handle very short social exchanges; ask and answer questions about what they do at school and in free time; make and respond to invitations; discuss what to do, where to go and make arrangements to meet; make and accept offers).
- Ability to perform simple classroom tasks in English.

Additional salient features of strong **Flyers (A2+)** candidates' performance:

- Ability to more actively participate in conversation.
- Ability to sustain monologues.

4. Which additional salient features can be identified in YLE candidate performances that are not currently covered in the CEFR?

A comprehensive list of salient features of children's reading abilities, activities, strategies and competences should be developed. Writing, as pointed out before, is not tested in YLE in any significant way. However, salient features of children's emerging writing abilities (copying, handwriting, spelling and other enabling skills) should also be developed for young learners along with the specific writing activities, strategies and competences children display in this age group. To identify and separately list salient features of typical candidate performance at each level of YLE that are *not* currently covered in the CEFR requires additional work on sample performances from YLE reading and writing papers as well as work with children in the classroom. Such work is being carried out in ECML research projects run by Angela Hasselgreen (personal communication, 2 June 2008).[4]

Recommendations

As a result of this exploratory study, it is recommended that salient features of YLE candidates' performance should be further developed, especially for children's reading and writing abilities, strategies and competences covering typical activities carried out in and outside the classroom to reflect YLE learners' language learning needs and language use.

An extension of this activity for Cambridge ESOL is the development of Can Do statements for YLE candidates (aged 7–12) to complement the list of Can Do statements already developed for 11–14 year-old school learners taking the KET and PET for Schools exams (Papp 2009). The Can Do statements are also expected to feed into work on formative assessment, also known as assessment for learning. The linking exercise generated further evidence of CEFR relatedness of the YLE suite as well as ideas for future reviews of the YLE tests.

Positive feedback from this exploratory study indicates that the primary objectives of the Council of Europe in encouraging exam boards to go through the alignment process for their exams, that is, awareness-raising of good testing practices and quality of tests claiming links to the CEFR, have been met in the case of the YLE tests. As a result of the exercise, it is possible to provide feedback to the Council of Europe on the linking process, for future revisions of the *Manual*, especially for relating exams such as the Cambridge YLE Tests to the framework of reference. Another spin-off of this linking exercise might be providing calibrated samples to the Council of Europe of children's performance in YLE tests as related to the CEFR.

However, the question remains whether it is necessary to carry out a whole-scale adaptation of current CEFR scales to reflect young learner needs and contexts or whether to start developing scales from scratch to reflect the nature of language learning and use by children, as pointed out by Papp (2007). Cambridge ESOL has followed both routes, first finding salient features shared by children and adults within the current version of the CEFR, particularly those for listening and speaking; and next developing scales and descriptors relevant for children's language learning needs and contexts, especially for reading and writing.

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

[4] European Centre for Modern Languages, Graz, Austria, see www.ecml.at

Cambridge ESOL's CEFR DVD of speaking performances: What's the story?

EVELINA GALACZI RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL
HANAN KHALIFA RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

This paper describes a process of developing a set of speaking test performances to exemplify a range of the Common European Framework of Reference (CEFR) levels. The newly developed selection of Cambridge ESOL speaking test performances coincides with the update of the Cambridge ESOL First Certificate of English (FCE) and Certificate in Advanced English (CAE) examinations in December 2008, the revision of the assessment scales for speaking for Main Suite and Business English Certificate (BEC) and the release of the final version of the Council of Europe *Manual* for relating language examinations to the CEFR (2009).

These Main Suite test selections could be used as calibrated samples in CEFR standardisation training and ultimately in aiding a common understanding of the CEFR levels. They can be found at Cambridge ESOL's website: www.cambridgeesol.org/what-we-do/research/speaking-performances.html. It should be noted that these are an additional resource to the existing speaking test samples on the Council of Europe's website, which are provided by Cambridge ESOL and Eurocentres, and the speaking performances recently compiled by the Centre international d'études pédagogiques (CIEP).[1] The following section explains the nature of Main Suite Speaking tests before moving on to provide a detailed description of the methodology used in selecting exemplar performances.

Cambridge ESOL's Main Suite Speaking Tests

The Cambridge approach to speaking is grounded in communicative competence models, including Bachman's (1990) Communicative Language Ability (which built on the work of Canale and Swain 1980 and Canale 1983) and the work of other researchers working in the field of task-based learning and assessment (Skehan 2001, Weir 2005b). As Taylor (2003) notes in her discussion of the Cambridge approach to speaking assessment, Cambridge ESOL tests reflect a view of speaking ability which involves multiple competencies (e.g. lexico-grammatical knowledge, phonological control, pragmatic awareness), to which has been added a more cognitive component which sees speaking ability as involving both a knowledge and a processing factor. The knowledge factor relates to a wide repertoire of lexis and grammar which allow flexible, appropriate, precise construction of utterances in real time. The processing factor involves a set of procedures for pronunciation, lexico-grammar and established phrasal 'chunks' of language which enable the candidate to conceive, formulate and articulate relevant responses with online planning reduced to acceptable amounts and timings (Levelt 1989).

In addition, Cambridge ESOL's approach to the assessment of speaking is based on a socio-cognitive model and an emphasis on the contextualisation of language use. Spoken language production is seen as situated social practice which involves reciprocal interaction with others, as being purposeful and goal-oriented within a specific context. Speaking, in other words, involves not just production, but also interaction, which is clearly reflected in the CEFR treatment of speaking as comprising two skills: production and interaction (Council of Europe 2001a:26). The main characteristic of a face-to-face speaking test is that interaction is bi-directional and jointly achieved by the participants, with the interlocutors accommodating their contributions to the evolving interaction.

Pairing of candidates where possible is a further feature of Cambridge ESOL Speaking tests which allows for a more varied sample of interaction, i.e. candidate-candidate as well as candidate-examiner. Similarly, the use of a multi-part test format allows for different patterns of spoken interaction to be produced, i.e. question and answer, uninterrupted long turn, discussion. The inclusion of a variety of task and response types is supported by numerous researchers who have made the case that multiple-task tests allow for a wider range of language to be elicited and so provide more evidence of the underlying abilities tested, i.e. the construct, and thereby contribute to the exam's fairness (Bygate 1988, Chalhoub-Deville 2001, Fulcher 1996, Shohamy 2000, Skehan 2001).

A further feature of the Cambridge ESOL Speaking tests is the authenticity of test content and tasks, as well as the authenticity of the candidate's interaction with that content (Bachman 1990). A concern for authenticity in the Cambridge ESOL exams can be seen in the fact that particular attention is given during the design and trialling stage to using tasks which reflect real-world usage, and are relevant to the contexts and purposes for use of the candidates.

Task specifications at all levels of the Speaking papers (e.g. in terms of purpose, audience, length, known assessment criteria, etc) are intended to reflect increasing demands on the candidate in terms of Levelt's (1989) four stages of speech processing. Tasks at the higher levels are more abstract and speculative than at lower levels and are intended to place greater demands on the candidates' cognitive resources. Scoring criteria are targeted at greater flexibility in the language used at the level of the utterance,

[1] See www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html and www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/videos/english.php

in interaction with other candidates or the examiner and in longer stretches of speech.

Cambridge ESOL's assessment scales

As well as informing Speaking test format and task design, the underlying construct of spoken language ability also shapes the choice and definition of assessment criteria, which cover Grammar/Vocabulary, Discourse Management, Pronunciation, and Interactive Communication. The use of both analytical and global criteria enables a focus on overall discourse performance as well as on specific features such as lexical range, grammatical accuracy and phonological control. (See KET, PET, FCE, CAE, CPE Handbooks for a more detailed explanation of assessment criteria.)

The Cambridge ESOL Main Suite Speaking scales span five global levels (KET/A2 to CPE/C2), which are in turn branched-out into sublevels (bands) in order to provide the possibility for a more fine-tuned dispersion of candidates taking Cambridge ESOL exams. Each global level in the Cambridge ESOL Speaking scale is broken down into 10 bands (Band 0, 1, 1.5, 2, … 4.5, 5). The categories in the scales are: Grammar and Vocabulary (Grammatical Resource and Lexical resource at levels C1 and C2), Discourse Management, Pronunciation, and Interactive Communication.

The descriptors for each level are stacked into a common scale, so that, for example, the descriptors at KET/A2 Band 5 are identical to those at PET/B1 Band 3 and FCE/B2 Band 1. This suggests some rough equivalencies between different bands for different levels. There are a few deviations from the 'stacking up' of levels: the descriptors for Pronunciation at levels C1 and C2 were identical, in line with current thinking on the assessment of Pronunciation (CEFR, Phonological Control Scale, Council of Europe 2001a:117), whereas the descriptors for Grammar and Vocabulary are worded somewhat differently in the transition from B2 to C1, since at C1 they are divided into two separate assessment criteria (Grammar/Vocabulary at A2–B2 and Grammatical Resource and Lexical Resource at C1–C2). Taking into account the overlap between some of the bands in the different levels, the result is a 25-point common scale covering levels A2–C2 (see Galaczi and ffrench 2007 for more detail on the development of the assessment scales).

Overview of procedures

In order to select oral performances exemplifying CEFR levels A2–C2, a project was set up. Project activities involved:

- the selection of oral performances from Cambridge ESOL's pool of video recorded speaking exams
- the identification and selection of raters
- a CEFR familiarisation exercise
- a marking exercise
- analysis and findings
- a collection of feedback on the use of the CEFR scales during the marking exercise.

Sample selection

Twenty eight test takers distributed in 14 pairs were selected for the purpose of the marking exercise. The test-taker samples came from a pool of existing Cambridge ESOL speaking test performances which are test recordings of speaking tests used for internal rater training purposes. In selecting the test takers to be used in the marking exercise, a range of proficiency levels and nationalities was targeted, and both male and female test takers were included.

The project consisted of two phases. Twenty test takers distributed in 10 pairs were used during Phase 1. They were taken from an available pool of 25 speaking tests which are employed for rater training purposes and are marked against the global and analytic Main Suite scales. The selection of the 10 pairs was based on the Main Suite marks awarded, and typical performances were operationalised as performances at the 3/3.5 band range of the Main Suite scale, while borderline performances were located at the 1.5/2 range of the scale. Based on the typical/borderline criteria adopted, one typical pair and one borderline pair were selected per level, to further confirm raters' ability to distinguish between borderline and typical candidates at different CEFR levels.

Phase 2 was an extension of the project and focused on performances at the two C levels only with a sample comprising four additional pairs of test takers (two at CAE and two at CPE). During this phase of the project a typical performance at CAE/C1 or CPE/C2 was operationalised as being at bands 4/4.5 of the Main Suite scale and a borderline performance was located at bands 2.5/3.

Entire speaking test performances, rather than test parts, were used in the sample in order to allow for longer stretches of candidate output to be rated by the raters. The use of whole tests also added a time-dimension to the project, as full tests are more time-consuming to watch and may introduce elements of fatigue. The raters had to spend a minimum of 8 minutes and a maximum of 19 minutes per single viewing. Such practical considerations limited the number of performances at each phase of the project to two per level.

Rater selection and profile

Eight raters participated in the project. They were chosen because of their extensive experience as raters for Main Suite speaking tests. They had also participated in previous Cambridge ESOL marking trials and had been shown through FACETS analyses to be within the norm for harshness/leniency and consistency. The raters had many years of experience as speaking examiners ranging from 11 to over 25 years, and were based in several parts of Europe. In addition, they had experience spanning different exams, with their corresponding task types and assessment scales, which had provided them with a richer and more in-depth experience as speaking examiners. In terms of familiarity with the CEFR, seven of the raters indicted that they were familiar/very familiar with the CEFR, while one rater reported a low-level of familiarity with the CEFR. As will be seen in the Instruments section below, a CEFR familiarisation activity given prior to the marking exercise was used to ensure that all raters had an adequate level of understanding of the CEFR before commncing rating the sample performances.

Research design and instruments

A fully-crossed design was employed where all raters marked all of the test performances on all of the assessment criteria. The decision to select 8 raters was based on recommendations given by Cizek and Bunch (2007:242). In addition, the number of observations recorded (8 raters giving 6 marks to 28 candidates) was in agreement with the sample size required by FACETS and allowed for measurements to be produced with a relatively small standard error of measurement.

The raters were sent the following materials:

- Two scales from the CEFR *Manual*: a global scale (Table C1, Council of Europe 2009a:184) and an analytic scale (Table C2, Council of Europe 2009a:185) comprising five criteria: Range, Accuracy, Fluency, Interaction, Coherence.
- A DVD with Main Suite Speaking tests (28 candidates total) arranged in random order.
- A CEFR familiarisation task.
- A rating form for recording the level awarded to each candidate and related comments.
- A feedback questionnaire.

The CEFR scales used were slightly adapted from the original, and 'plus levels' were added across the whole scale. It was felt that the raters needed to have the full-range of the scale available, so they could make finer distinctions between the levels of the speaking performances, including A1+ and C1+, which are not in the global and analytic CEFR speaking scales (Council of Europe 2009a:184–5). Taking into account the borderline 'plus' levels, the scales used in the project had 12 points.

CEFR familiarisation activity and a marking exercise

The raters were sent detailed instructions about the marking, which are given in Figure 1 in the next column. Steps 1 through 3 aimed to familiarise or refresh raters' understanding of the CEFR scales for oral assessment and to establish a common interpretation of the descriptors.

Data analysis

The marks awarded by the raters and the responses to the feedback questionnaire were compiled in an Excel spreadsheet. The marks were further exported into SPSS to allow for the calculation of descriptive statistics and frequencies. In addition, a Multi-Faceted Rasch analysis was carried out using the program FACETS. Candidate, rater, and criterion were treated as facets in an overall model. FACETS provided indicators of the consistency of the rater judgements and their relative harshness or leniency, as well as fair average scores for all candidates.

Figure 1: Raters' instructions for marking

Please go through the following steps:

1. Read through the CEFR scales to get a feel for the detail of description for the global and analytic categories (Range, Accuracy, Fluency, Interaction, Coherence).
2. Highlight key elements of the descriptors that indicate differences in performance at each level.
3. Do a self-assessment exercise in order to become more familiar with the scales prior to rating. Think of a foreign language you speak. If you do not speak a foreign language, think of a specific language learner who you have taught in the past or a language learner you are familiar with. Assess that learner using the global assessment scales first. Then give an assessment for each of the categories in the analytic scales. Record your ratings on the form given.
4. Start rating the candidates on the DVD. Assess each test in the order given on the DVD.
5. To make an assessment, start with the global assessment scale in order to decide approximately what level you think the speaker is. Assign a global rating during your first 2–3 minutes of the test. Then change to the analytic scales and assess the candidates on all five criteria (Range, Accuracy, Fluency, Interaction, Coherence). As you are watching, note features of candidate output to help you arrive at your final rating and refer to the scales throughout the test.
6. At the end of each test, enter your marks for each assessment criterion on the rating form. Add comments to explain your choice of marks, linking your comments to the wording of the CEFR descriptors, and giving examples of relevant candidate output where possible. You may need to watch the test again to cite examples but your assessments should not be changed. Please limit the number of viewings of each test to a maximum of two.
7. NOTE: Even if you can recognise the test from the materials used (e.g., KET, PET, etc.), it is important not to assign a CEFR level automatically, based on your prior knowledge of the test. Use the descriptors in the CEFR scales, so that you provide an independent rating, and support your choice of level by referring to the CEF.
8. Complete the feedback questionnaire.

Findings

Ascertaining the consistency and severity of the raters was an important first step in the analysis, as it gave scoring validity evidence to the marks they had awarded. The FACETS output generated indices of rater harshness or leniency and consistency. As seen in Table 1, the results indicated a very small difference in rater severity (spanning 0.37 to -0.56 logits), which was well within an acceptable severity range and gave no cases of unacceptable fit (all outfit mean squares were within the 0.5 to 1.5 range), indicating high levels of examiner consistency. These results signalled a high level of homogeneity in the marking of the test, and provided scoring validity evidence (Weir 2005b) to the ratings awarded.

Table 1: FACETS output: Rater severity and consistency

Rater	Measure (logit)	Standard Error	Outfit MnSq
1	.37	.09	.62
2	-.24	.10	.80
3	.35	.09	1.32
4	-.19	.10	.70
5	.31	.09	1.10
6	-.20	.10	.78
7	-.56	.10	.95
8	.16	.09	1.17

Phase 1 results

The results indicated very strong rater agreement in terms of typical and borderline performances at levels A2 to B2. As noted earlier, the internal team's operationalisation during sample selection had considered a performance at Main Suite band 3/3.5 as typical of a given level and a performance at Main Suite band 1.5/2 as borderline. This operationalisation had worked very well at levels A2 to B2 and the selection of performances which the internal group had felt to be typical/borderline (based on the marks awarded against the Main Suite scale) was confirmed by the high agreement among the raters in assigning CEFR levels across all assessment criteria to those performances.

At levels C1 and C2 there was a lower level of agreement among raters regarding the level of the performances; in addition, the marking produced mostly candidates with differing proficiency profiles and so no pair emerged as comprising two typical candidates across all assessment criteria at the respective level, which led to extending the project into a second phase. The lower degree of agreement among raters at the higher proficiency levels was most likely because it is simply more difficult to mark higher-level candidates whose output is more complex and therefore leaves more room for divergent evaluations.

The raters' marks for each performance at the C levels also resulted in a CEFR level which was consistently lower than what was predicted by the Main Suite mark. It is not possible to be certain why the discrepancy between Main Suite and CEFR C levels occurred. One possibility is that any test format inevitably imposes constraints on the quantity and quality of candidate language produced and as such a typical, solid C2 test performance may not necessarily match the C2 level performance descriptors which are designed to cover a much broader context.

We can also hypothesise that the CEFR C-levels and the corresponding Main Suite CAE/CPE levels have developed more independently than the lower levels. The CEFR and the Cambridge levels have a common origin in Wilkins (1976) and are the result of a 'policy of convergence' (Brian North 2007, personal communication) and an 'interactive process of evolution' (Nick Saville 2006, personal communication). The historical and conceptual relationship between the CEFR and Cambridge ESOL scales indicates that the work on the Waystage, Threshold and Vantage levels (van Ek and Trim 1998b, 1998a, 2001) seems to have progressed very much hand-in-hand between the Council of Europe and Cambridge ESOL (Taylor and Jones 2006), and so a 'tight' relationship there is to be expected. In addition, the C levels have not been described in same detail as the lower levels, which have already been addressed by the Waystage, Threshold and Vantage specifications. Perhaps it is worth mentioning here that establishing criterial features at the C levels is one of the major aims of the English Profile Programme (see www.englishprofile.org, also Salamoura and Saville 2009 in this issue). It should be noted that Milanovic made a similar point in his opening article of this issue (see Milanovic 2009).

The lower level of agreement among raters regarding candidates at C1 and C2, and the difficulty of finding a pair of candidates typical of these two levels across all criteria introduced the need for a subsequent marking exercise which focused on the top two levels only. The Phase 1 result led to a change in the group's working operationalisation of a typical C1 and C2 performance as measured against the Main Suite scale. As such, performances in the 4/4.5 band range were selected for the subsequent Phase 2 of the study.

Phase 2 results

The results from this phase produced a typical pair of test takers at C1 level across all CEFR assessment criteria, with very high rater agreement. The pairs used at C2 had more varied performances and no pair emerged as having two typical C2 performances across all assessment criteria. This result is not altogether surprising given that the performances used in the present exercise came from the rater training pool where both typical and borderline cases should feature to allow for raters to develop familiarity with a range of test taker abilities. The C2 pair which was selected, therefore, included one typical candidate at that level across all criteria, while the second test taker in the pair showed borderline performance at the C1/C1+ level.

The selection of the final sample

Taking the statistical evidence into account five pairs of candidates emerged as the most suitable Main Suite illustrations for levels A2 to C2, as shown in Table 2.

It is important to note that in compiling this selection of Speaking tests, we have made our best effort to select typical performances. However, we would like to draw the reader's attention to the fact that educational contexts and traditions vary from one country to another and this may

Table 2: Selected performances

Candidate	Overall level	Range	Accuracy	Fluency	Inter-action	Coherence
Mansour	**A2**	A2	A2	A2	A2	A2
Arvids	**A2**	A2	A2	A2	A2	A2
Veronica	**B1**	B1	B1	B1	B1	B1
Melisa	**B1**	B1	B1	B1	B1	B1
Rino	**B2**	B2	B1+/B2	B2	B2	B2/B2+
Gabriela	**B2**	B2	B2	B2	B2	B2
Christian	**C1**	C1	C1	C1	C1	C1
Laurent	**C1**	C1	C1	C1	C1	C1
Ben	**C1/C1+**	C1	C1	C1/C1+	C1+	C1
Aliser	**C2**	C2	C2	C2	C2	C2

have an effect on perceptions of typical levels of performances. For example, our experience in international benchmarking projects has indicated that in certain educational contexts aspects of fluency are more favoured than aspects of accuracy and vice versa.

After the selection of performances was finalised, commentaries were provided for each selected performance, which included positive comments about what this learner Can Do, as well as an explanation of why they are not at the level above. An example from a B1 level candidate is given in Table 3, and the full set of commentaries can be found online.[2]

Table 3: Sample candidate commentary

VERONICA: LEVEL B1

Veronica can link phrases into connected speech to produce simple but comprehensible language. She is reasonably accurate in predictable situations, and so demonstrates a B1 level of performance. She does not produce the level of detailed description or accuracy required in a B2 performance.

Range (B1):
Veronica has sufficient range to express herself on familiar topics *('I have a collection of dolls … typical dress the countries'. 'My favourite time is holidays and summer because I can go to the beach and enjoy the sun and the sea with my friends and with my parents')*, although there is some hesitation searching for language *('her mother … er … cook … er … delicious cake', 'I prefer give a gift for a woman flowers … er … or …. er … card … small card … chocolate')*.

Accuracy (B1):
Veronica's language is reasonably accurate, enabling her to express her intended meaning, although there are frequent non-impeding errors throughout *('I am study diplomatic science, because I be studying to be an ambassador' 'there are a grandmother.')*. Meaning depends on context on one occasion *('It's important the food, he needs to eat someone …')*.

Fluency (B1):
Veronica is able to string sentences and phrases together, although with evident pausing to organise the language *('This is a Japanese family … there are a grandmother, grandfather, because … it's the birthday of a child it's all the family together … her mother cooked a delicious cake and on the table are some gifts for the birthday. All the family are very happy')*.

Interaction (B1):
Interaction is appropriate throughout the test. In Part 2 Veronica responds appropriately to her partner and introduces her own ideas *'yeah, sure, and …', 'maybe, …, I think it's more important this moment the guitar.'* She maintains the discussion on Part 4, picking up appropriately on the interlocutor's prompt.

Coherence (B1):
Veronica uses simple connectors to link phrases coherently *('I prefer something special, no expensive but something special … maybe flowers is OK for me maybe chocolate', 'I have something that is is very good for me: if someone give me is a dolls because I have a collection of dolls er with different kinds of typical dress er the different countries, and for me is very good gift')*.

Rater feedback on using the CEFR descriptors

In addition to providing examples of speaking performances at a range of CEFR levels, this project also provided an opportunity to explore the raters' experience of using the CEFR global and analytic oral scales. At the conclusion of the marking exercise, raters were asked to evaluate their ease of applying the CEFR global oral assessment scale (Council of Europe 2009a:184) and the more analytic scale (Council of Europe 2009a:185). A six-point scale was used for evaluation where 1 = Very Easy and 6 = Challenging. With regards to the ease of use of the global scale, 6 raters chose 'challenging' (categories 4, 5 or 6), while 2 raters opted for 'easy' (1, 2, 3). In terms of the ease of application of the analytic scales, 3 raters chose 'challenging', and 5 raters preferred 'easy'.

The raters also responded to a question on the ease of application of the assessment criteria, and their responses are given in Table 4 below. With such a small number of raters, it is difficult to reach any definite conclusions, but some trends can nevertheless be discerned, such as the choice of 'Fluency' as an easier criterion to apply, in contrast to 'Range' and 'Coherence' as more difficult criteria for the raters. The feedback below explores some of the reasons behind these choices.

Table 4: Raters' perception of ease of applying the CEFR analytic scales

Easiest criterion (no. raters)	Most difficult criterion (no. raters)
Fluency (4)	Range (3)
Interaction (2)	Coherence (3)
Accuracy (2)	Interaction (1)
	Fluency (1)

The common themes running through the feedback are given below. This feedback gives an insight into the application of the scales to test performances.

Applying the CEFR Global Assessment scale

One of the features of the CEFR global scale is that it asks raters to apply it during the first 2–3 minutes of a performance, so they can arrive at an approximate decision on a candidate's overall level. Many raters commented on the difficulty of awarding a reliable global score based on just a couple of minutes of interaction, especially at the higher levels:

- Very few interviews give a broad enough sample in the first 2–3 minutes – especially at higher levels.
- Using the scale in the first 2–3 mins of a test is not so easy sometimes, as the candidates usually don't say anything very specific in this phase of the test (introductions and personal questions). It was quite hard to differentiate at the top 3 levels early on in a test. The summary lines for each level were very useful as a reminder and the paragraphs were good reminders of the main points of each level.
- Especially at the higher levels, I didn't feel I had enough sample from the first few minutes to be able to give a reliable global mark. I think the long turn is important to me when giving a global mark.
- The wording of the scales, especially as you go further up, means that it is hard to apply them with any degree of precision to the sample of language produced by a candidate during the first 2–3 minutes of a Cambridge exam – the descriptor for B2 as an example, refers to 'a wide range of topics', 'stretches of language', 'clear, detailed descriptions on a wide range of subjects'; you

[2] See www.cambridgeesol.org/what-we-do/research/speaking-performances.html

just don't get these in part 1 of an FCE exam, which is at the B2 level.

- Clearly, the initial part of each test covers familiar topics and candidates appear at time [sic] to be deceptively fluent. It was quite difficult to award a realistic score on this basis.
- Applying the scales themselves was not so much the problem; it was applying them in the first three minutes which was difficult, as the sample of language was not always sufficient; e.g. 'clear, detailed descriptions of complex subjects' at C1 is unlikely to take place in the first three minutes.

Assessment of Pronunciation

A strong theme running through the rater feedback focused on the need for a Pronunciation assessment criterion, especially at the lower levels:

- When pronunciation impeded understanding and so judgement of accuracy/range etc., it was not clear where to penalise this.
- I also felt there was a need for a mention of pronunciation features somewhere at the lower levels, where unclear pronunciation can affect comprehensibility.
- I also found myself distracted by having nowhere to 'deal with' assessment of pronunciation.
- Scales lacked a specific category for pronunciation and occasionally stress and intonation affected meaning/interaction.
- Something I missed, probably because of the mindset I have developed during my years as an examiner using the Cambridge criteria, was any reference to pronunciation. I tried hard not to let it influence me, as it wasn't present in the criteria, but couldn't help noticing that some candidates had pronunciation which helped their cause, while others were hampered by theirs.

The lack of Pronunciation descriptors in the analytic and global CEFR scales is discussed in North and Hughes (2003:5), who note that the production of pronunciation descriptors was found to be problematic since 'pronunciation tends to be perceived as a negative phenomenon, interference from mother tongue, rather than as a positive competence. This makes it difficult to scale mathematically with positive concepts.' The authors go on to add that 'it is actually extremely difficult to define a set of ascending levels of pronunciation ability [and] learners with the same background and the same language level can vary wildly in their pronunciation' and conclude that 'the fact of the matter is that one cannot have the same confidence in [the 'Phonological Control'] scale as in the scales for other aspects of spoken language (2003:6). Therefore it is not included in the criteria grid.' It is worth noting that the feedback from the raters in this project also agrees with the general feeling among the L2 assessment community (as seen in the recent discussion on the Language Testing discussion list L-TestL, May 2009) that the inclusion of descriptors for the assessment of pronunciation would enhance the main CEFR scales.

Assessment of Accuracy/Range

The raters in this study commented extensively on these two assessment criteria, and there was some consensus that the 'Range' criterion was the most difficult to apply:

- The accuracy scales seemed to give too much emphasis to candidates correcting their own mistakes (B2 and C1), which is not something candidates tend to do naturally, I don't think, in an exam. Maybe there should be more emphasis on 'errors which occur when attempting more complex language', rather than 'errors are rare' for the top levels.
- There was a small part of the descriptor which made the rest of it difficult eg: C1 Accuracy – the last comment about 'errors being generally corrected when they do occur' made applying this descriptor difficult.
- I found myself distracted by thinking about the wording of the Range and Accuracy descriptors – I'd like clearer references to lexical knowledge.
- My feeling is that the Accuracy and Range descriptors seem to be weighted towards structural more than lexical knowledge: I missed the more explicit references to lexis that we have in Cambridge descriptors, especially at the higher levels.
- Range covers a lot of ground (vocabulary, sentence patterns, expressions, colloquialisms, ability to describe/explain/reformulate/emphasise etc) and the language of the descriptors becomes increasingly complicated as the levels progress.
- The 'range' criterion was the hardest to apply because of the limited amount of data provided during the course of the test. For example, the CPE candidate Ryon[3] may well have possessed a greater range than he was able to display during the test, but the limitations of the situation and time available, plus his own rather reserved manner, made it impossible to judge this.

Assessment of Coherence

The raters noted some difficulty in applying this criterion at the lower levels:

- Also coherence at the lower levels seems to be totally dependent on linkers, which is rather limited I think. (A1, A2 and B1 are hardly any different from each other) Not sure coherence is necessarily evident in these terms at the very low levels.
- Although criteria [sic] is 'coherence', most of the descriptors relate to cohesion – especially in the lower levels.
- I found the description of coherence at the lower levels difficult to apply in practice.

Assessment of Fluency

There was a general consensus among the raters that the Fluency descriptors were easy to apply:

- I like the use of a 'fluency' descriptor, and the notions of 'flow', 'ease' and 'effort'.

[3] Name has been changed.

- I like the references to 'ease', 'flow', 'smooth flow', 'effortless', 'pre-packaged utterances' etc. I sometimes find that as assessor during live tests that there doesn't seem to be anywhere obvious to deal with these aspects.
- The differences in the levels was quite clear for fluency.

Assessment of Interaction

The raters also provided some comments on the close relationship between the 'Interaction' criterion and task type. A couple of them commented on the resulting difficulty of assessing interaction with certain task types at the lower levels, where the tasks are more controlled.

- Descriptions sometimes did not fit format of examinations e.g. A2 (KET) Interaction – no 'conversation' is required – suggest change to 'interaction'. C1 Interaction – nothing about turn taking. Predictable format/language requirements of topics in lower level exams makes it difficult to judge whether strong candidates at that level could cope with more general topics and therefore fulfil the criteria for a higher level.
- Some of the descriptors seemed to bring features together which perhaps do not sit together very well e.g. B1 and B2 Interaction refer to topics as well as interaction skills.

Conclusion

This small-scale project has focused on just one aspect of a much larger endeavour. The selection of speaking performances to exemplify different CEFR levels is a necessary step in providing support for stakeholders in interpreting different proficiency levels. A natural extension of this project is to investigate the comparability between the two scales used here (CEFR and Main Suite) from several perspectives, and to focus on the comparability between the assessment criteria, the performance descriptors, and the bands. These issues will be explored in further studies and reported in future issues of *Research Notes*.

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

Some evidence to support the alignment of an LSP Writing test to the CEFR

HUGH BATEMAN ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

It is widely accepted within the language testing community that it is not enough simply to make claims of test validity, but rather that it is necessary to provide evidence. In practice, language testers have in fact been slow to publish evidence of this nature (Weir 2005b), although two recent studies, Shaw and Weir (2007) and Khalifa and Weir (2009), represent important contributions to the evidence-based validity literature.

Since the publication of the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001a), concern has grown that insufficiently validated claims are being made of alignment with the CEFR, with the danger of confusion among test users, and a potential loss of credibility for the CEFR (Figueras et al. 2005, Weir 2005a). In response to this concern, Weir (2005b) set out to describe 'the basic minimum requirements for sound testing practice. Stakeholders in the testing process, in particular students and teachers, need to be able to ask the right questions of any examinations' (2005b:12). In order for these questions to be asked and answered, Weir constructed frameworks for the validation of tests of reading, writing, listening and speaking, drawing on socio-cognitive models of language use in those skill areas.

This article describes a study that applied one of these frameworks to investigate an aspect of the context validity of a Cambridge ESOL test, namely the linguistic demands it made of the candidates who took it.

The examination investigated in the study was a test of Language for Specific Purposes (LSP), BEC Vantage. Designed to measure at CEFR Level B2, BEC Vantage is the most widely taken of the Business English Certificate (BEC) exams in English for business purposes. Figueras et al. (2005) point out the advantages of mapping tests to the CEFR at the level of each language skill as well as for the overall assessment, and this study chose to focus on the Writing skill. Weir (2005b) suggests, while recognising that this may not be possible operationally, that each test version should be validated individually. The study therefore investigated a single version, the May 2006[1] test, which was sat by around 20,000 candidates in 20 countries.

Context validity

Weir (2005b:19) proposes the term context validity, which adds 'the social dimension of language use' to the traditional notion of content validity. This has echoes in Douglas (2000), who describes the situational context of language use as 'the social, physical and temporal situation the language activity is taking place in' (ibid.:42) and notes that: 'a context is not simply a collection of features imposed upon a language learner/user, but rather is constructed by the participants in the communicative event' (ibid.:43).

[1] Although the test was sat on 3 June 2006, it belonged to the May exam session.

Context validity also draws upon Bachman's notion of situational authenticity, defined as 'the perceived relevance of the test method characteristics to the features of a specific target language use situation' (Bachman 1991:690). For Bachman, a test task with high situational authenticity would have the same 'critical features' (ibid.:691) as tasks in the domain of language use that it sought to measure. With his socio-cognitive framework for test validation, Weir (2005b) sets about identifying these critical features, not only in terms of the task itself, but also the conditions of its administration and the linguistic demands it makes on candidates.

The framework appears particularly useful in the validation of tests of LSP, such as BEC, since the importance of context has long been a concern of LSP testing. The notion that language use depends on context (Douglas 2000) provides a useful theoretical foundation for producing tests of LSP. Situational authenticity has also proved to be a useful concept in 'justifying' LSP testing (Douglas 2000, Dudley-Evans and St John 1998).

O'Sullivan (2006) argues against an LSP/non-LSP dichotomy, instead seeing all language tests as being placed somewhere on a continuum of specificity, and conceptualises specificity in terms of context validity (as well as in terms of the cognitive processing that tasks elicit).

What, then, is the context of language use that BEC aims to measure candidates' ability to function within? The *BEC Vantage Information for Candidates* booklet states that:

> 'The Business English Certificates (BEC) from Cambridge ESOL have been created specifically for individual learners who wish to obtain a business-related English language qualification. By taking an internationally recognised business qualification like BEC, you can show that you have learned English to an appropriate standard and can use it in a professional context.' (UCLES 2006:2)

To focus specifically on context in relation to writing tasks, in recent decades interest has grown in the social aspects of writing activity (Weigle 2002). Traditionally, writing has been thought of as the result of cognitive effort on the part of the writer. Increasingly, however, it is seen as a social and cultural act: 'What we write, how we write, and who we write to is shaped by social convention and by our history of social interaction' (Hayes 1996:5).

With regard to writing in the workplace, individual professions are characterised by 'a unique set of cognitive needs, social conditions and relationships with society at large' (Gunnarsson 1997:5 cited in Bargiela-Chiappini and Nickerson 1999:1), and the culture of individual organisations may shape, and be shaped by, norms regarding, for example, whether certain types of communication are generally written or spoken, and the level of formality required in intra-organisational written communication. Figure 1 lists some of the aspects of context validity for testing writing (Shaw and Weir 2007).

Figure 1: Aspects of context validity for writing (Shaw and Weir 2007:64)

Context validity

Setting: task
- Response format
- Purpose
- Knowledge of criteria
- Weighting
- Text length
- Time constraints
- Writer–reader relationship

Setting: administration
- Physical conditions
- Uniformity of administration
- Security

Linguistic demands:
Task input and output
- Lexical resources
- Structural resources
- Discourse mode
- Functional resources
- Context knowledge

Linguistic demands: task input and output

For a test task to have high context validity, the linguistic demands that must be met for successful task realisation should be appropriate. These linguistic demands need to be as similar as possible to those made by equivalent tasks in real-life language use at the level of performance we are targeting if we are to generalise from test performance to language use in the future domain of interest (Shaw and Weir 2007:91).

Unfortunately there appears to be little published research into the linguistic demands of real-life business writing tasks in either the LSP or business communication literature. However, attempts to compare the linguistic demands of the test with the literature relating to level B2 of the CEFR were rather more successful.

Lexical resources

The CEFR (Council of Europe 2001a) is intended to apply to many languages, and does not attempt to specify the actual lexical items in any language that characterise its levels. (However, see Salamoura and Saville 2009 for an overview of the ongoing work of the English Profile Programme to develop a wordlist for English covering levels A1 to B2.)

The CEFR gives the following summary of vocabulary range at B2:

> 'Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.' (Council of Europe 2001a:112)

It also summarises vocabulary control at B2: 'Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication' (ibid.).

While noting the increasing individualisation and divergence of learners' lexical development after B1 level, van Ek and Trim (2001) suggest that 'in order to carry out the tasks described for Vantage adequately, the learners will need to have a command of vocabulary that allows them to express themselves precisely and with some subtlety […] while being sensitive to shades of meaning, implications and overtones' (van Ek and Trim 2001:77–8).

In the absence of detailed guidance on the lexical resources consistent with task input and candidate output at B2 level in a work context, further evidence that the test complies would necessarily be indirect, and, for reasons of space, does not form part of the study reported here. Future research could perhaps make use of learner corpora to investigate whether the lexical resources required by the task are consistent with B2 level, following lexical studies in other contexts (see Barker 2008, Horner and Strutt 2004, Rose 2008, Wright 2008).

However, there is evidence that appropriate steps were taken to ensure that the lexical input in every test version would be familiar to the candidates who typically enter the exam (some of whom have not reached the required language level and fail the examination). Item writers are trained in order to give them a high awareness of the level, and the tasks that they produce are edited by a team before being content vetted, and then pretested. Moving on to the lexis required to complete the task, tasks are constructed so that successful completion of the task does not depend on the inclusion of specific items of lexis, including business-specific lexis, that do not appear in the input.

See figures 2 and 3 for the Writing tasks answered in the BEC Vantage test studied. The Part 1 task in May 2006 was pretested in October/November 2004 on 66 students preparing for BEC Vantage at exam centres in Argentina, China, Italy, Poland and the UK. Twenty three of the students also chose to complete a feedback questionnaire which among other things required them to agree with one of the following statements in relation to the Part 1 task: ‘The instructions were clear’; ‘The instructions were not clear’. All 23 students agreed with the statement ‘The instructions were clear’. The Part 2 task was pretested between March and June 2004 on 43 students preparing for BEC Vantage at exam centres in Argentina, Austria, China, France and Switzerland. Eleven students chose to complete the feedback questionnaire, all of them agreeing with the statement ‘The instructions were clear’.

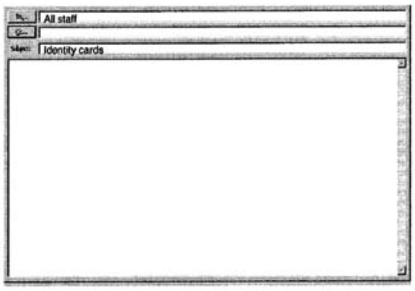
Structural resources

Similar endeavours are made to ensure that the task is accessible to the target candidature in terms of the structural resources required to interpret the task.

Figure 2: May 2006 BEC Vantage Test of Writing Part 1

PART ONE

- The software company you work for has decided to introduce identity cards for certain staff in your department.
- Write an **e-mail** to all staff in your department:
  - saying which staff will need identity cards
  - explaining why the identity cards are needed
  - informing staff how to get a card.
- Write **40–50** words.
- Write on the opposite page.

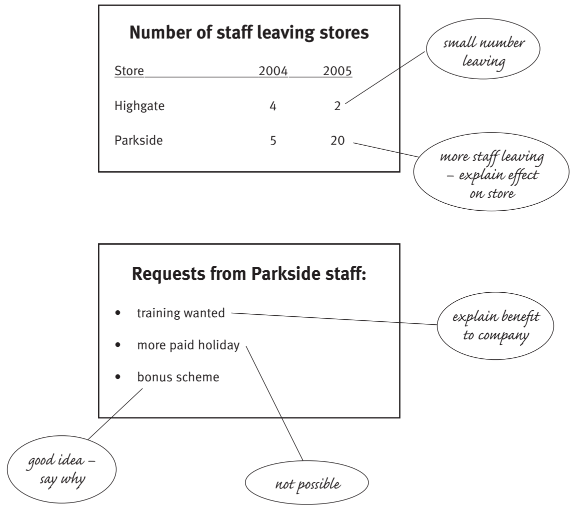

All staff

Identity cards

Figure 3: May 2006 BEC Vantage Test of Writing Part 2

PART TWO

- The number of staff leaving Parkside, one of your company's retail stores, is high compared with another of its stores. Your line manager has asked you to write a report about the situation.
- Look at the information below, on which you have already made some handwritten notes.
- Then, using **all** your handwritten notes, write your **report**.
- Write **120–140** words.
- Write on the opposite page.


Number of staff leaving stores

Store	2004	2005
Highgate	4	2
Parkside	5	20

small number leaving

more staff leaving – explain effect on store

Requests from Parkside staff:

- training wanted
- more paid holiday
- bonus scheme

explain benefit to company

good idea – say why

not possible

The structures used in the task instructions and stimulus material are deliberately kept within the limits that characterise the structural competence of B2 learners, for example by limiting the use of longer sentences containing subordinate clauses (on the May 2006 paper there are three), notwithstanding the fact that: ‘At Vantage Level learners will be able to follow and produce longer discourses structured by such means as [...] the construction of complex sentences by the embedding of subordinate clauses’ (van Ek and Trim 2001:3–4).

The CEFR does not give advice on the structures that candidates might be expected to use at each level. However, van Ek and Trim (1998a) provide a detailed grammatical summary of level B1, and although an equivalent summary in relation to B2 does not appear in van Ek and Trim (2001), it is made clear that the step up from B1 to B2 is not characterised by the addition of large numbers of new structures. Indeed, van Ek and Trim's (1998a) grammatical summary of level B1 suggests that B1 learners are likely to have learnt the structures needed to successfully complete the test, particularly as business language is not known for its use of elaborate structures.

The May 2006 paper does, therefore, seem fair to candidates in terms of the structural resources required of them. Research into the actual structures used by the candidates could be carried out by analysing the responses to each task stored in the CLC (for related work, see Barker 2008, Horner and Strutt 2004).

Discourse mode

For several reasons, it is problematic to provide evidence that the discourse modes required in the test tasks are consistent with CEFR level B2. First, as Shaw and Weir note:

> 'Investigating the nature and impact of discourse mode is [...] beset by two problems. First, there is little agreement in the literature on the terminology that should be used to classify different texts and second, the effect of texts required on the difficulty level of the task is not that well researched at the moment.' (Shaw and Weir 2007:115)

Second, perhaps because of the lack of agreement on terminology and paucity of research, the CEFR does not attempt to specify which written discourse modes characterise the levels.

In her 'dimensions of tasks for direct writing assessment', Weigle (2002) lists three dimensions that might be considered constituent parts of discourse mode:

> 'The genre refers to the expected form and communicative function of the written product; for example, a letter, an essay, or a laboratory report. The rhetorical task is broadly defined as one of the traditional discourse modes of narration, description, exposition, and argument/persuasion, as specified in the prompt, while the pattern of exposition [...] refers to subcategories of exposition or specific instructions to the test taker to make comparisons, outline causes and effects, and so on.' (Weigle 2002:62)

The study chose to focus on the first of these, genre.

Van Ek and Trim (2001:93) characterise writing at this level as follows: 'The learners can perform, within the limits of the resources available to them at vantage level, those writing tasks which adult citizens in general may wish, or be called upon, to carry out in their private capacity or as members of the general public'.

However, they devote little over a page to writing genres, and make no mention of writing for business or work purposes. Consequently, there is no indication of the suitability of a report, the genre required in Part 2 of the test, as a writing task at this level; nor is there any mention of the genre required in Part 1, email, as the publication originally appeared in the 1990s.

Although there is little evidence that genres required by the test are consistent with B2, there is rather more evidence that they are consistent with real-life writing in the workplace. The use of email 'has come to prominence in the modern workplace as a major element in business information retrieval and use' (Mulholland 1999:57), cutting across business sector and to some extent across the role and even the position in the hierarchy of the writer. It is widely used by managers among others: Nickerson (1999) reports Markus' 1994 study of managers in a large corporation, for whom email was 'the primary medium of internal work-related communication'.[2] Furthermore, 'understanding and writing faxes, letters, memos, email' is one of the activities listed for the work context by the Association of Language Testers in Europe's Can Do statements (ALTE 2002:80). It therefore seems highly appropriate that one of the tasks on the BEC Vantage May 2006 paper should be to write an email.

In order to widen the domain of target language use that the exam covers, the genre required in Part 2 varies from one exam session to another, and may be a piece of correspondence (a letter, fax or external email), a report or a proposal. The Part 2 task on the May 2006 paper was to write a report. This would appear appropriate given that 'understanding and writing reports (of substantial length and formality)' is one of the activities listed for the work context by the ALTE Can Do statements (ALTE 2002:80). A literature review by St John (1996) finds that corresponding and report writing are the two writing activities at the core of business communication skills, and the fact that business skills courses for both L1 and L2 students commonly include report writing suggest that it is a required skill in the real-life world of work. In their review of writing for professional purposes, Grabe and Kaplan (1996) state that internal reports, progress reports and project proposals are important genres in workplaces, albeit less common than letters, memos, forms and instructions.

Functional resources

In the Part 1 task, the candidate is required to give information, and explain. These functions seem appropriate for a workplace email writing task, given that 'email is primarily used to exchange information in organisational settings' (Nickerson 1999:40). Nickerson's assertion is supported by Sherblom (1988:49), who, in a study of 157 emails received by a middle-level computer services manager in a large organisation, found that: 'mail designed to exchange information was sent more frequently than was mail involving more complex communication functions such as personal, social, and influence attempts'.

In the Part 2 task, the candidate is required to give information, evaluate it, and make recommendations. Once again, no published research was found into the functional resources required by real-life report writing tasks in either the LSP or business communication literature. However, anecdotal evidence suggests that these are common functions of reports in the real-life workplace.

Regarding the question of whether the functional resources required by the task are consistent with B2, van Ek and Trim (2001:7) list the language functions that B2 learners will need to perform, noting that 'there is no fundamental difference' between these and the B1 equivalent. The functions required by both test tasks appear to map well to van Ek and Trim's categories 'Imparting and seeking information' and 'Deciding on and managing courses of action: suasion', suggesting that test candidates' functional resources were not unfairly stretched by the test tasks. However, it should be noted that van Ek and Trim's list does not appear to actively consider written discourse (all exemplification is in the form of spoken discourse) or workplace communication.

The Council of Europe (2001a) provides illustrative scales for language use at each CEFR level, both for overall written production (see Figure 4) and for the sub-categories of 'creative writing' and 'reports and essays'. These scales were created by recombining elements of descriptors from other scales and 'have not been empirically calibrated with the measurement model' (ibid.61). Nevertheless, the scales seem to support the claim that the linguistic demands of the May 2006 BEC Vantage Writing Part 2 task was appropriate for candidates at CEFR level B2.

The task requires candidates to synthesise and evaluate information, which are B2 descriptors, but not to write on

[2] Markus, L 1994, Electronic mail as the medium of managerial choice, *Organisation Science* 5:502–27, quoted by Nickerson 1999:38.

Figure 4: Overall written production (Council of Europe 2001:61)

	Overall written production
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.
C1	Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.
B2	Can write clear, detailed texts in a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.
B1	Can write straightforward, connected texts on a range of familiar subjects within his field of interest by linking a series of shorter, discrete elements into a linear sequence.
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.
A1	Can write simple isolated phrases and sentences.

'complex subjects' or to support points of view 'at some length with subsidiary points, reasons and relevant examples', which are C1 descriptors.

The reference in the B1 descriptor to 'straightforward connected texts [...] linking a series of shorter discrete elements into a linear sequence' seems to describe the Part 1 task fairly well. However, it is likely that subject matter of the Part 1 task will be merely *related* to the candidate's field of interest, a B2 descriptor, rather than a *familiar subject* within his or her field of interest, as specified for B1.

Although the Council of Europe does not provide more detailed descriptors for a task such as Part 1, its sub-scale for 'reports and essays' is relevant to the Part 2 task, shown in Figure 5.

Figure 5: Reports and essays (Council of Europe 2001:62)

	Reports and Essays
C2	Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
C1	Can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support points of view at some length with subsidiary points, reasons and relevant examples.
B2	Can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem.
	Can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
B1	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated, factual information on familiar, routine and non-routine matters within his/her field with some confidence.
	Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
A2	No descriptor available
A1	No descriptor available

The lower of the two B2 descriptors, which one assumes corresponds to the lower level of B2 performance (B2.1) described on pages 33 and 35 of the CEFR (Council of Europe 2001a), appears to be most consistent with the demands of the Part 2 task. In particular, the final three content points require candidates to 'give reasons in support of or against a particular point of view and explain the advantages and disadvantages of various options'.

More detailed than the scales referred to above are those developed and validated by ALTE (2002), and both tasks appear to correspond to CEFR level B2 in the ALTE Can Do statements that relate to writing for work purposes.

Under the activity which best describes the Part 1 task, i.e. 'Understanding and writing faxes, letters, memos, e-mail, etc.', the Can Do statement for CEFR level B2 is 'CAN write a non-routine letter where this is restricted to matters of fact' (ALTE 2002:78).[3] Following the argument above, an answer that remains at the level of fact will be acceptable, but an answer above the level of the test may bring in persuasion. The Can Do statement for CEFR level C1, which is the next level up, begins 'CAN write most letters (s)he is likely to be asked to do' (ibid.).

Under the activity which best describes the Part 2 task, i.e. 'Understanding and writing reports (of substantial length and formality)', the Can Do statement for CEFR level B2 is 'CAN write a simple report of a factual nature and begin to evaluate, advise etc' (ALTE 2002:80). The five content points in the May 2006 Part 2 task do progress from being factual at the start of the task to evaluating and advising as the task progresses.

In conclusion, then, there is evidence that the functional resources required in the test tasks are consistent both with real-life workplace emails and reports and with CEFR level B2.

Content knowledge

In order to complete the May 2006 tasks, as well as language knowledge, candidates will need a degree of content knowledge, in the form of a general, basic knowledge or understanding of how things are likely to operate in the world of work.

In Part 1, this knowledge probably entails: being able to suggest a category of staff in a software company who might need security badges; being able to suggest a reason for needing security badges; and being able to suggest a procedure for obtaining a security badge. In Part 2, it is likely to entail: being able to suggest what effect high staff turnover would have on a store; being able to suggest a benefit to the company associated with training its staff; and being able to suggest a reason that a bonus scheme would be a good idea.

From a traditional view of (general purposes) language testing, this is problematic, since it is important that 'the test score reflects the area(s) of language ability we want to measure, and very little else.' (Bachman and Palmer 1996:21). However, although a need for content knowledge would not be appropriate for a test of language for general

[3] Above CEFR level A2, the statements for this activity invariably refer to letters rather than faxes, memos or email.

purposes, it is not necessarily inappropriate for a test of English language for business purposes such as BEC Vantage (Douglas 2000, O'Sullivan 2006):

> 'The interaction between language knowledge and content, or background, knowledge is perhaps the clearest defining feature of LSP testing, for in more general purpose language testing, the factor of background knowledge is usually seen as a confounding variable, contributing to measurement error and to be minimized as much as possible. In LSP testing […] background knowledge is a necessary, integral part of the concept of specific purpose language ability.' (Douglas 2000:2)

The context validity of the test is therefore enhanced rather than threatened, because people performing similar real-life writing tasks are likely to possess, and indeed require, content knowledge of the type required by the test task: 'There needs to be a congruence between the types of knowledge and tasks the test requires and the types of knowledge and tasks the test demanded by the situation for which the tests results are to be interpreted, the target language use situation' (Douglas 2002:30).

Underpinning the theoretical justification, pretesting confirmed that the content knowledge required to perform each task did not systematically exceed that possessed by BEC Vantage candidates.

Conclusion

This study provides the following evidence to support the alignment of an LSP test of Writing, the May 2006 BEC Vantage Writing test, to Level B2 of the CEFR in terms of the linguistic demands of the tasks.

The test production process ensured that the task input did not overextend the lexical resources of the target candidature, that successful completion of the tasks did not depend on the use of specific items of lexis, and that the task was accessible to the target candidature in terms of the structural resources required to access the task and respond to it. The discourse mode of each test task was consistent with writing tasks in the real-life workplace, and the functional resources required by the test tasks were consistent both with the requirements of real-life workplace emails and reports, and with CEFR Level B2. Finally, appropriate steps were taken to ensure that the content knowledge required to access and complete each task did not exceed that possessed by the target candidature.

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

Criterial features of English across the CEFR levels: evidence from the English Profile Programme

ANGELIKI SALAMOURA RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL
NICK SAVILLE RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

English Profile (EP) and criterial features

The English Profile is a collaborative programme of interdisciplinary research, whose goal is to provide a set of Reference Level Descriptions (RLDs) for English for all six levels of the Common European Framework of Reference (CEFR) from A1 to C2 (Council of Europe 2001a). See Little 2007 for an extended discussion of the CEFR.

The EP website provides an overview of the EP research programme (www.englishprofile.org; see also Kurtes and Saville 2008, and Salamoura 2008 in a special issue of *Research Notes*, August 2008). A main focus of EP research is the identification of 'criterial features' of English for each CEFR level, or in other words, how each level differs from adjacent levels (cf. Hendriks 2008). This paper summarises the approach and outcomes to date.

Towards a definition of criterial features

A 'criterial feature' is one whose use varies according to the level achieved and thus can serve as a basis for the estimation of a language learner's proficiency level. So far the various EP research strands have identified the following kinds of linguistic feature whose use or non-use, accuracy of use or frequency of use may be criterial: lexical/semantic, morpho-syntactic/syntactic, functional, notional, discourse, and pragmatic. A more detailed inventory will be produced as the research progresses.

What makes a feature 'criterial' is an open question which the EP researchers have been addressing as part of their collaborative agenda. In fact, the programme has adopted an iterative approach to formulating and testing research questions and hypotheses: as empirical evidence is accumulated and shared, more criterial features will be identified. The more the criterial features are understood in relation to the empirical data, so the research questions will be refined over time.

In the *Corpus and Computational Linguistics* strand, Hawkins and Buttery (2009) have identified four types of feature that may be criterial for distinguishing one CEFR level from the others. Although couched primarily in grammatical terms (i.e. lexical semantic, morpho-syntactic and syntactic features), this classification may also be extended to encompass other types of language feature. The four categories are described below.

1. Acquired/Learnt language features

These are language features a learner masters at a given level and uses them accurately and consistently at the higher levels. In this category fall the '*positive grammatical properties*' that Hawkins and Buttery (2009:2) describe as:

> '…correct properties of English that are acquired at a certain L2 level and that generally persist at all higher levels. E.g. property P acquired at B2 may differentiate [B2, C1 and C2] from [A1, A2 and B1] and will be criterial for the former. Criteriality characterises a set of adjacent levels in this case. Alternatively some property might be attained only at C2 and be unique to this highest level.'

2. Developing language features

These are features that appear at a certain level but they are unstable, i.e. they are not used correctly in a consistent way. This category includes what that Hawkins and Buttery (2009:2) call '*negative grammatical properties*' of an L2 level, i.e.:

> '…incorrect properties or errors that occur at a certain level or levels, and with a characteristic frequency. Both the presence versus absence of the errors, and the characteristic frequency of the error (the 'error bandwidth') can be criterial for the given level or levels. E.g. error property P with a characteristic frequency F may be criterial for [B1 and B2]; error property P with frequency F' may be criterial for [C1 and C2].'

Hawkins and Buttery (2009:7) define criteriality for 'negative grammatical properties', i.e. errors, as follows:

> 'An error distribution is criterial for a level L if the frequency of errors at L differs significantly from their frequency at the next higher and lower levels, if any. Significance amounts to a difference of at least 29% from level to level, which guarantees at least one standard deviation from the mean. Two or more levels can be grouped together for criteriality if each is not significantly differentiated from any immediately higher and lower levels (i.e. by less than 29%).'

Given the evolving nature of second language acquisition and learning, one would predict that several language features would pass through a developing stage before they are acquired or learnt. So, one feature that is still developing at one proficiency level may be acquired at the next level up, or a feature may be developing across more than one level.

3. Acquired/Native-like usage distributions of a correct feature

Hawkins and Buttery (2009:2) describe acquired or native-like usage as follows:

> 'Positive usage distributions for a correct property of L2 that match the distribution of native speaking (i.e. L1) users of the L2. The positive usage distribution may be acquired at a certain level and will generally persist at all higher levels and be criterial for the relevant levels, e.g. [C1 and C2].'

4. Developing/Non native-like usage distributions of a correct feature

The final category – developing or non native-like usage is described by Hawkins and Buttery (2009:2) as:

> 'Negative usage distributions for a correct property of L2 that do not match the distribution of native speaking (i.e. L1) users of the L2. The negative usage distribution may occur at a certain level or levels with a characteristic frequency F and be criterial for the relevant level(s), e.g. [B2].'

Criterial features of English across the CEFR levels

In this section, we provide a 'snapshot' of the learner data gathered and research conducted to date and outline some preliminary findings with regards to the criterial features identified so far. As noted, it is expected that these findings will be refined, revised and complemented as more data become available and as more research is carried out.

The research findings of Filipovič (2009), Hawkins and Buttery (2009), Hendriks (2008), Parodi (2008) and Williams (2007) outlined below are based on the *Cambridge Learner Corpus* (CLC) which has been parsed using the RASP parser (Briscoe, Carroll and Watson 2006).[1] The findings related to *language functions* outlined by Green (2008) are derived from a database of materials at B2, C1 and C2 levels collected from a variety of countries worldwide, including test specifications, proficiency scales, syllabuses, course book outlines and English language curricula. This database was compiled at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, UK.

Acquired language features and native-like usage distributions of a correct feature

B1 level

At B1 level there are a number of language features that have been acquired, or show native-like usage. These are listed below.

New *verb co-occurrence frames* are reported by Williams (2007) and are listed in Table 1 (reproduced from Hawkins and Buttery 2009).

For Spanish and German learners of English only (Parodi 2008): Subject-verb agreement, the syntax of questions and negation for both main verbs with lexical content (*go*, *arrive*, *walk*, *drive*, *cycle*) and modal and auxiliary verbs (*will*, *can*, *must*, *have*, *be*, *do*). Other cross-linguistic B1 level criterial features are:

- Inflection for person or not: I walk – he walk vs. I can – he* cans
- Inflection for tense or not: I walk – I walked vs. I must
- Finite and non finite forms: I cycle, I have cycled, I can cycle; I expect to cycle, I expect to have arrived; I expect to *can/be able to cycle
- Questions: can/will Kim read? what does Kim read? *what reads Kim?
- Negations: Kim can/will/does not drive, *Kim drives not.

B2 level

As for B1 level, there are new verb co-occurrence frames for B2 level reported by Williams (2007), shown in Table 2 (reproduced from Hawkins and Buttery 2009).

[1] See the Cambridge Learner Corpus webpages: www.cambridge.org/elt/corpus/learner_corpus.htm

Table 1: New verb co-occurrence frames at B1 level

Frame	Example
NP-V-NP-NP	*She asked him* [*his name*]
NP-V-Part	*She gave up*
NP-V-VPinfin (WH-move)	He explained [*how to do it*]
NP-V-NP-V(+*ing*) (Obj Control)	*I caught him stealing*
NP-V-NP-PP (P=*to*) (Subtype: Dative Movement)	*He gave* [*a big kiss*] [*to his mother*]
NP-V-NP-(*to be*)-NP (Subj to Obj Raising)	*I found him (to be) a good doctor*
NP-V-NP-Vpastparti (V=passive) (Obj Control)	*He wanted* [*the children*] *found*
NP-V-P-V*ing*-NP (V=+*ing*) (Subj Control)	*They failed in attempting the climb*
NP-V-Part-NP-PP	*I separated out* [*the three boys*] [*from the crowd*]
NP-V-NP-Part-PP	*I separated* [*the three boys*] *out* [*from the crowd*]
NP-V-S (Wh-move)	*He asked* [*how she did it*]
NP-V-PP-S	*They admitted* [*to the authorities*] [*that they had entered illegally*]
NP-V-S (*whether* = Wh-move)	*He asked* [*whether he should come*]
NP-V-P-S (*whether* = Wh-move)	*He thought about* [*whether he wanted to go*]

Table 2: New verb co-occurrence frames at B2 level

Frame	Example
NP-V-NP-AdjP (Obj Control)	*He painted* [*the car*] *red*
NP-V-NP-*as*-NP (Obj Control)	*I sent him as* [*a messenger*]
NP-V-NP-S	*He told* [*the audience*] [*that he was leaving*]
NP-V-P-NP-V(+*ing*) (Obj Control)	*They worried about him drinking*
NP-V-VPinfin (Wh-move) (Subj Control)	*He thought about* [*what to do*]
NP-V-S (Wh-move)	*He asked* [*what he should do*]
NP-V-Part-VPinfin (Subj Control)	*He set out to win*

C1 level

At Cl level, relative clauses in genitive positions (*the professor whose book I read*) seem to be criterial as is the relative distribution of indirect object/oblique relative clauses (*the professor that I gave the book to*) compared to relative clauses on other positions (subjects, direct objects and genitives) (Hawkins and Buttery 2009).

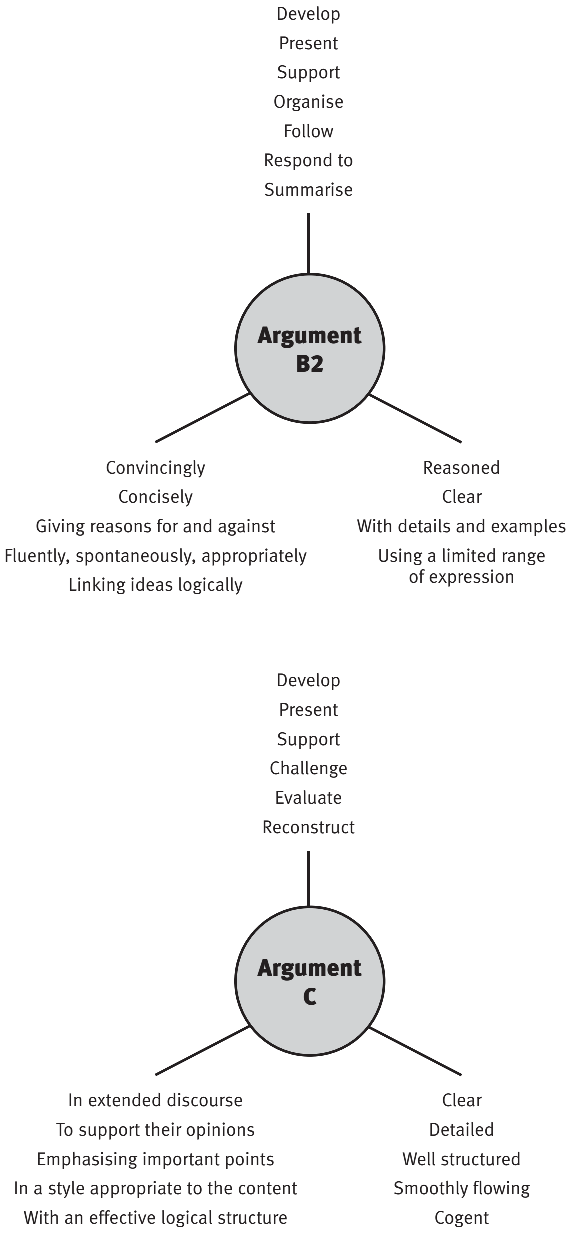
C2 level

At C2 level, the relative distribution of subject relatives (*the professor who wrote the book*) to object relatives (*the book that the professor wrote*) is a strong candidate for a criterial feature (Hawkins and Buttery 2009:3)

The C levels (C1 and C2) compared with B2

There are some functional features which become prominent at the C levels. For example, *argument*, *suasion* and *rational enquiry and exposition* increase at C levels in comparison to B2 level (see Green 2008, who considered Wilkins' 1976 list of functions and the Can Do statements in the CRELLA database). These functions may be found at B2 level, but Green's analysis suggests a shift in focus in their use at the C levels which makes them particularly relevant to the C levels. An example is the word *inform* (expressing *argument*); synonyms such as *tell* or *give information* occur at B2 level but the choice of *inform* at the C level 'may suggest rather a subtle change of perspective' (Green 2008:22). Figure 1 below, which shows the linguistic contexts of use for *argue* (and its derivatives *arguing*, *argues*, *argument*, *arguments*, *argumentative*) at the B2 and C levels, is a good illustration of this 'subtle change' from B2 to C levels.

Figure 1: Contexts for *argument* and its derivatives at B2 and C levels (Green 2008)


Develop
Present
Support
Organise
Follow
Respond to
Summarise
Argument B2
Convincingly
Concisely
Giving reasons for and against
Fluently, spontaneously, appropriately
Linking ideas logically
Reasoned
Clear
With details and examples
Using a limited range of expression
Develop
Present
Support
Challenge
Evaluate
Reconstruct
Argument C
In extended discourse
To support their opinions
Emphasising important points
In a style appropriate to the content
With an effective logical structure
Clear
Detailed
Well structured
Smoothly flowing
Cogent

The keywords: *coherent*, *colloquial*, *confidently*, *conveys*, *critically*, *demonstrates*, *edit*, *evaluate*, *finer*, *genre*, *integrating*, *interprets*, *proficiency*, *slang*, *structurally*, *structured*, *subtleties* were found to be unique at the C levels (Green 2008).

Developing language features and usage distributions of a correct feature

In this section we outline criterial features and distributions of features whose *error rates* show that the accuracy of their use is developing over a number of CEFR levels. The symbol '›' used below means 'significantly higher error rate than'.

Progressive error patterns in B1–C2

These are declining errors from B1 to C2 levels (after Hawkins and Buttery's Table 6, 2009:8–9).

B1 › B2 › C1 › C2

- Derivation of determiner (*She name was Anna*)
- Form of determiner (*I have an car*)

B1 › B2 › [C1 and C2]

- Incorrect inflection of verb (*I spended last weekend in London*)

[B1 and B2] › C1 › C2

- Wrong tense of verbs (*I spend last weekend in London*)
- Missing preposition (*I gave it John*)
- Replace quantifier (*It all happened a lot of years ago*)
- Argument structure error (*It gives great pleasure to me*)

[B1 and B2] › [C1 and C2]

- Missing quantifier (*I'll call in the next days*)

[B1 and B2 and C1] › C2

- Replace verb (*I existed last weekend in London*)
- Derivation of verb error (*I spendified last weekend in London*)
- Replace adverb (*He stared at her intensively*)
- Derivation of adverb error (*It happened fastly*)
- Replace preposition (*When I arrived at London*)
- Quantifier countability error (*It cost him many money*).

Inverted U error patterns in B1–C2 [B]

These are errors that increase after B1 (over B2 and C1) and decline again at C2 (after Hawkins and Buttery's Table 7, 2009:9–10).

B2 › [B1 and C1] › C2

- Noun agreement error (*One of my friend*)
- Unnecessary verb (*I spend to be last weekend in London*)
- Missing verb (*I last weekend in London*)
- Complex error (*He didn't never should be having*)

B2 › C1 › [B1 and C2]

- Missing noun (*It was an interesting*)
- Countability error (*I don't have any monies*)

[B2 and C1] › B1 › C2

- Verb agreement error (*The three birds is singing*)

[B2 and C1] › [B1 and C2]

- Unnecessary determiner (*There was a lot of the traffic*)
- Replace noun (*Have a good travel*)

B2 › [B1 and C1 and C2]

- Missing conjunction (*The stripes were red green*)
- Determiner agreement error (*I enjoy these job*).

Common verbs

Common verbs such as *know*, *see*, *think*, *want*, *get*, *go*, *say*, *come*, *need* are overrepresented in the learner output in comparison to native speaker usage. This overrepresentation declines at the higher levels (Hawkins and Buttery 2009:10–11).

Indirect object/oblique relative clauses

The relative distribution of indirect object/oblique relative clauses (*the professor that I gave the book to*) compared to relative clauses on other positions (subjects, direct objects and genitives) departs at A2, B1 and B2 levels from that of native speakers of English (Hawkins and Buttery 2009).

Verbs expressing spatial information

Verbs expressing spatial information (e.g. path or manner of movement or manner of attachment) have been categorised into verbs that are more or less neutral in spatial information (1st tier verbs, e.g. *put*, *go*, *come*, *take*) and verbs that express specific information (2nd and 3rd tier verbs, e.g. skip, hop, pierce). Hendriks (2008) found a tendency of decreasing proportions of 1st tier (more neutral) verbs from the A2 to the C2 level, and increasing proportions of 2nd and 3rd tier (more specific) verbs.

L1-specific criterial features

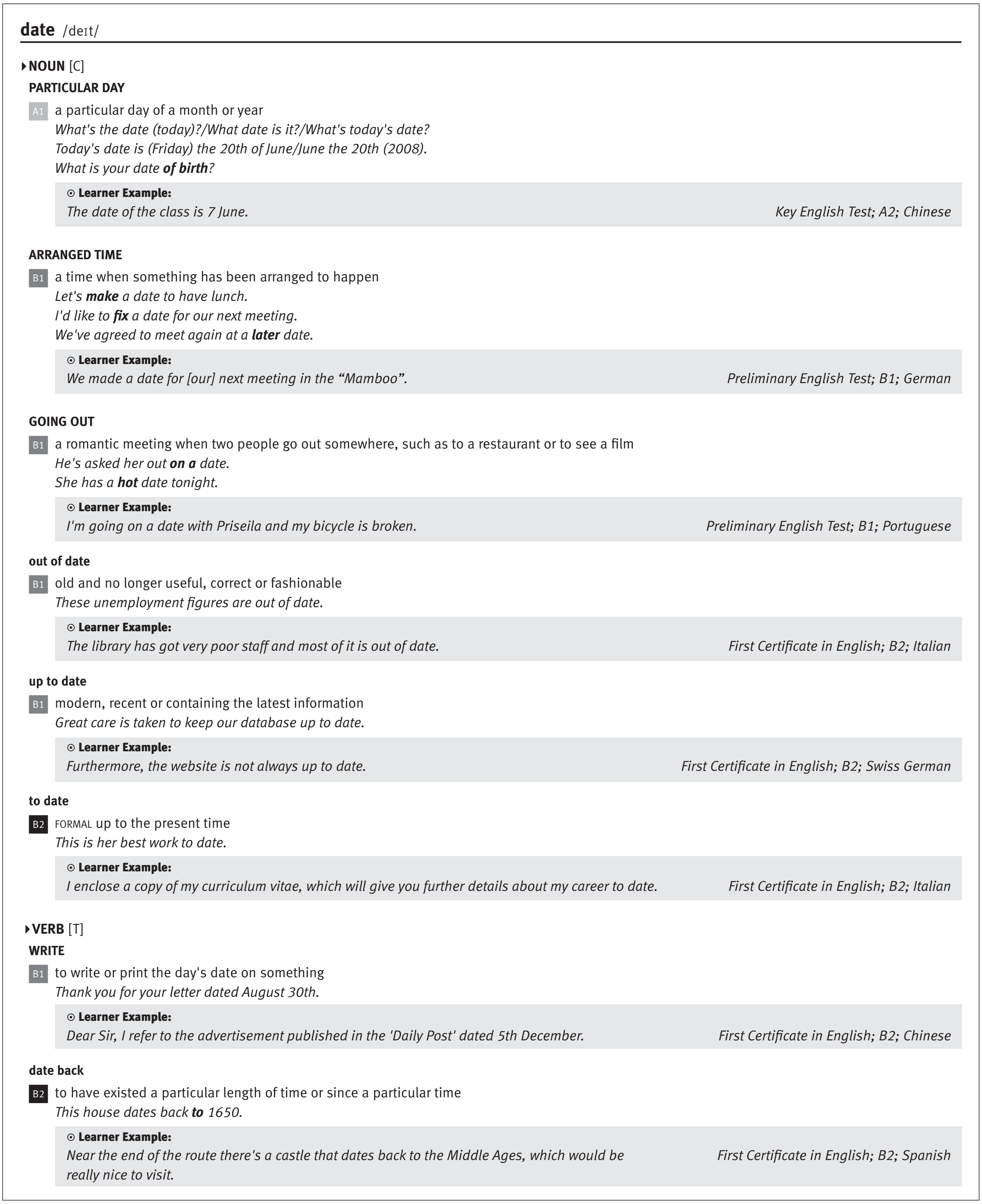
Speakers of languages without definite and indefinite articles have significantly higher rates of missing determiner errors in L2 English across B1–C2 than speakers of languages with articles (Hawkins and Buttery 2009:11).

Regarding *spatial verbs* expressing voluntary motion only, Hendriks (2008) hypothesised that '[if] transfer from the L1 were to occur, one would expect Spanish learners of English to produce less Manner and more Path verbs, whereas German and Chinese speakers should encode more Manner and Cause in the verbs [like native English speakers would do]'. Hendriks found that whereas Chinese learners of English follow the expected pattern in that they produce Manner verbs from A2 to C2, German speakers, contrary to expectation, do not show this developmental path (e.g. no Manner verbs occur at A2 level at all. However this latter finding should be interpreted with caution due to the very small number of spatial verbs in the German sample studied. Contrary to expectation again, Spanish learners of English do start using Manner verbs from very early on, i.e. A2. These results cannot be simply explained by transfer from the mother tongue. Therefore, Hendriks (2008) suggests that one possible explanation may be that at the earliest proficiency levels, the preferred way of expressing voluntary motion with verbs, if there is a problem encoding semantic information, is Path rather than Manner or Cause irrespective of the learners' first language.

English Profile Wordlists

The Wordlist strand of EP research focuses specifically on vocabulary development across the levels (Capel 2009). The first output of this research is the EP Wordlists, a comprehensive listing of words and phrases in English that are considered to be within the CEFR levels starting with the first four levels: A1–B2. The Wordlists provide information at word and sense level, based on extensive analysis of word frequency and learner use, using the Cambridge International Corpus, the British National Corpus and the Cambridge Learner Corpus, together with other sources, including the Cambridge ESOL vocabulary

Table 3: The word 'date' – an example from English Profile Wordlists



date /deɪt/

▸**NOUN** [C]

PARTICULAR DAY

A1 a particular day of a month or year
What's the date (today)?/What date is it?/What's today's date?
Today's date is (Friday) the 20th of June/June the 20th (2008).
*What is your date **of birth**?*

⊙ **Learner Example:**
The date of the class is 7 June. — *Key English Test; A2; Chinese*

ARRANGED TIME

B1 a time when something has been arranged to happen
*Let's **make** a date to have lunch.*
*I'd like to **fix** a date for our next meeting.*
*We've agreed to meet again at a **later** date.*

⊙ **Learner Example:**
We made a date for [our] next meeting in the "Mamboo". — *Preliminary English Test; B1; German*

GOING OUT

B1 a romantic meeting when two people go out somewhere, such as to a restaurant or to see a film
*He's asked her out **on a** date.*
*She has a **hot** date tonight.*

⊙ **Learner Example:**
I'm going on a date with Priseila and my bicycle is broken. — *Preliminary English Test; B1; Portuguese*

out of date

B1 old and no longer useful, correct or fashionable
These unemployment figures are out of date.

⊙ **Learner Example:**
The library has got very poor staff and most of it is out of date. — *First Certificate in English; B2; Italian*

up to date

B1 modern, recent or containing the latest information
Great care is taken to keep our database up to date.

⊙ **Learner Example:**
Furthermore, the website is not always up to date. — *First Certificate in English; B2; Swiss German*

to date

B2 FORMAL up to the present time
This is her best work to date.

⊙ **Learner Example:**
I enclose a copy of my curriculum vitae, which will give you further details about my career to date. — *First Certificate in English; B2; Italian*

▸**VERB** [T]

WRITE

B1 to write or print the day's date on something
Thank you for your letter dated August 30th.

⊙ **Learner Example:**
Dear Sir, I refer to the advertisement published in the 'Daily Post' dated 5th December. — *First Certificate in English; B2; Chinese*

date back

B2 to have existed a particular length of time or since a particular time
*This house dates back **to** 1650.*

⊙ **Learner Example:**
Near the end of the route there's a castle that dates back to the Middle Ages, which would be really nice to visit. — *First Certificate in English; B2; Spanish*

lists and classroom materials targeted at the different levels.

The Wordlists could thus be considered 'criterial vocabulary knowledge' for each CEFR level (A1–B2). A preview version of the EP Wordlists (letters D, J and K) is now available on the EP website. Table 3 provides an example from the Wordlists.

The T-series revisited

Researchers within the EP team are currently revisiting the T-series publications (*Breakthrough*, Trim 2001b, *Waystage*, *Threshold* and *Vantage*, van Ek and Trim 1998b, 1998a, 2001) in search of features that are novel at each level and that could thus qualify for the status of criterial features (Filipovič 2009, Hawkins in preparation). The criteriality or not of these features will be checked and confirmed against the EP empirical learner data. Tables 4 and 5 provide some preliminary results from this project.

Future directions

These initial findings form a good test bed for refining the initial hypotheses and for pointing to future research directions. A number of the questions that EP researchers are currently addressing include:

- How do the different kinds of criterial features (lexical semantic, morpho-syntactic, syntactic, discourse, notional, functional, etc.) interrelate? Answering this question is fundamental in bringing together the different strands of EP research.

Table 4: Novel language features by part of speech across A2–B2 sourced from the T-series (after Appendix 1 in Filipovič 2009)

Parts of speech	Waystage A2	Threshold B1	Vantage B2
Nouns		• Plurals-different pronunciation of endings • Nouns in singular only • Genitive 's – different pronunciation	• Compound nouns and complex N and NP genitives
Pronouns		• Reflexive/emphatic (*myself*, etc.) • Gender in 3rd person singular • Anaphoric use for non-sex-specific personal nouns (*he/him/their*, etc.)	
Determiners		• *the*-different pronunciation; differentiating uses: unique, generic • *a/an*-frequency (once a day), amount (two pounds a kilo) • Identifiers: *other/another*	
Adjectives		• V+ing (*walking stick*) • V+ed/en (*broken promises*) • Attributive only (*daily*) vs. predicative only (*alive*) • Gradable (e.g. polar *old* vs. *young*) and non-gradable (*married/single*) • Comparison of gradable Adjs (*such*, *like*, *the same*) • Equality/inequality: *as…as/not so…as; different from* • Complementising Adjs: broader spectrum than A2	
Adverbs	• Interrogative uses: who, how, etc. • *Where is your pen? I don't know.*	• Relative uses: *Where is your pen? I do not know where it is.* • Preference: *rather than*	
Prepositions	• Simple prepositions in transparent uses	• Complex prepositional phrases: *in the centre of*, *in the neighbourhood of*, *to the left/right of*, etc.	
Verbs	• *may*-permission • *must*-withholding permission • *should*-advice • *will*-future reference, requests, intentions	• Present Simple for future reference with adverbs: *The train leaves soon*. • Past Perfect: all uses • Present Cont. with future reference: *We are driving to Scotland next week*. • Past Cont.: all uses except reporting • Present Perfect Cont. and Past Perfect Cont.: all uses • More complex passives: *A book was given to me/I was given the book*. • *may*-possibility • *might*-all uses, e.g. suggesting a course of action • *must*-necessity, logical necessity, pressing invitation • *should*-duty, expectation • *will*-prediction, capacity	
Conjunctions		• *as well as* • *as strong as* • Effect, consequence: *It was so hot that/so I took my coat off.* • Relative: *I know what you mean.*	• General vantage point: word formation and compounding; prefixes: *anti-, de-, dis-, non-, pre-, re-, un-,* suffixes: *-able, -hood, -ify, -less, -like, -ness*

Table 5: Novel language features at the phrase and clause level across A2–B2 sourced from the T-series (after Appendix 1 in Filipovič 2009)

Phrase and clause level	Waystage A2	Threshold B1	Vantage B2
VP		• +benefactive: *I gave John the letter for Mason*. • +instrumental: *Susan opened the door with a key*. • to+INF as subject: *To kill people is wrong*. • Following complementising adjectives and verbs: *He is likely/ expected to arrive late*.; *He forgot to lock the door*. • Gerund as subject: *Swimming is good for you*.	• Phrasal verbs:V+adverbial particle (+NP): *What did the wind blow down? The wind blew down the tree*. vs. *What did the wind blow down?* *The wind blew (very hard) down the valley.*
ADV P	*We eat in the kitchen*.	• Increased complexity: *We drove to the seaside by car*. • Equality/Inequality: *He did as well as he could*.	
Pronoun P.		• PRO+Adjunct: *May I have something to drink?* • Indefinite PRO+Adj.: *He told me nothing new.* • Indef, PRO+relative clause: *Susan is someone I met in Spain*.	
Adj P.		• Predicative Adj.+postmodifier: *This food is not good enough*. • Predicative Adj.+Adjunct: *Smoking is not good for you*. • Predicative complementising: *Apples are good to eat*.	• Adverbs of degree+gradable Adjectives: *She is a very beautiful and most intelligent woman*. • Adjectives+past participle: *This is a very poorly made dress*.
Clause level		• Adjectival and adverbial relative clauses • Following It+certain complementising verbs, adjectives and NPs: *It does not matter that she is not there. It is likely that it will snow tonight. It is a pity that they cannot come*. • WH+NP+VP (as subject and following be complement)&WHAT+VP (as subject and object): *What I like is watching TV*. *This is not what I wanted*. *What interests me most is politics*. *I know what is meant.* • NP+to+VPinf. *I want my son to be a doctor.* • NP+VPgerund: *I remember my brother being born.* • NP+to be+Adj.: *I prefer water to be boiled.* • NP+VPinf. *I saw him drive away.* *I had the laundry clean my coat.*	• Nesting of further clauses and phrases

- What lexical semantic, morpho-syntactic, syntactic and discourse features are exponents of notions and functions (notional and functional features) and vice versa?
- To what extent does the criteriality of features vary depending on the L1 of the learner?
- What is the effect of task type on learner production and criterial features? (Parodi 2008)
- How does the type of context in which spatial verbs occur help explain the spatial information findings? (Hendriks 2008)
- Which language functions relate to the needs of younger learners and special groups such as adult migrant learners of English? (Green 2008)
- How does cognitive complexity interact with linguistic complexity in younger learners? (Green 2008)
- Are Can Do statements useful tools for language description at the C levels or do we need to find alternative ways to describe English language functions at the higher levels? (Green 2008)
- How are the language functions operationalised by course and test providers? An extended analysis of the content of textbooks and test materials to which they relate may shed light into this issue (Green 2008).
- Does EP entail the same kinds of empirical validation across all its research areas?

The immediate future of the EP will involve extending the current analyses to broader samples from the CLC and collecting other kinds of written data from learners of English worldwide. Another major challenge being addressed is how to include *spoken language* in the analysis (McCarthy and Saville 2009), as well as other data that will make it possible to foster a closer relationship between the EPP outcomes and teachers and learners of English in their different contexts worldwide (Alexopoulou 2008).

Conclusion

It is envisaged that the description of English across the CEFR levels in terms of criterial features will result in a valuable data source for researchers and a useful tool for practitioners in the fields of English language learning, teaching and assessment.

Moreover, as an outcome of the EP, it is hoped that the CEFR itself can be operationalised more effectively for English and that it will become a more useful tool for its intended purposes. The search for criterial features will lead to better *linguistic descriptions*, and this in turn will lead to better *functional descriptors*, thus addressing a current weakness (see Milanovic 2009 in this issue). Already the focus on empirical research at the bottom and top ends of the scale (A1, and C1/2) is providing more precise information about the nature of proficiency in English at these levels. As the English Profile Programme continues, more evidence will come to light about the nature of language proficiency at all levels of the CEFR.

References and further reading

For a full bibliography from this issue, including these references, see pages 41–44.

Bibliography on CEFR and Cambridge ESOL

This bibliography contains all of the sources from issue 37 August 2009, our special issue describing Cambridge ESOL's approach to, and use of, the CEFR. It should be referred to alongside individual articles. A list of all previous *Research Notes* articles related to the CEFR is given on page 44.

Alderson, J C (Ed.) (2002) *Common European Framework of Reference for Languages: learning, teaching, assessment – Case studies*, Strasbourg: Council of Europe, available online www.coe.int/T/DG4/Portfolio/documents/case_studies_CEF.doc

— (2007) The CEFR and the Need for More Research, *Modern Language Journal* 91/4, 659–63.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2006) Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of the Dutch CEFR Construct Project, *Language Assessment Quarterly*, 3/1, 3–30.

Alexopoulou, T (2008) Building new corpora for English Profile, *Research Notes* 33, 15–19, Cambridge: Cambridge ESOL.

ALTE (2002) *The ALTE Can Do Project* (English version), available online www.alte.org/downloads/index.php?doctypeid=10

— (2005) CEFR *Grid for the Analysis of Speaking Tasks*. Version 2.0, available online www.coe.int/T/DG4/Portfolio/documents/ALTE CEFR Speaking Grid OUTput51.pdf

Andrich, D (1978) Relationships between the Thurstone and Rasch approaches to item Scaling, *Applied Psychological Measurement* 2, 449–460.

Bachman, L F (1990) *Fundamental considerations in language testing*, Oxford: Oxford University Press.

— (1991) What does language testing have to offer? *TESOL Quarterly* 25/4, 671–704.

Bachman, L F, Davidson, F and Milanovic, M (1996) The use of test method characteristics in the content analysis and design of EFL proficiency tests, *Language Testing* 13, 125–150.

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*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.

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

Bargiela-Chiappini, F and Nickerson, C (Eds) (1999) *Writing Business: Genres, Media and Discourses*, London: Longman.

Barker, F (2008) Exploring lexical differences in General English Reading papers, *Research Notes* 31, 22–26, Cambridge: Cambridge ESOL.

Bateman, H (2009) Some evidence supporting the alignment of an LSP Writing test to the CEFR, *Research Notes* 37, 29–34, Cambridge: Cambridge ESOL.

Bramley, T (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment, *Journal of Applied Measurement* 6/2, 202–223.

Briscoe, E, Carroll J and Watson R (2006) *The second release of the RASP System*, in Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia, available online http://acl.ldc.upenn.edu/P/P06/P06–4020.pdf

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

Cameron, L (2001) *Teaching Languages to Young Learners*, Cambridge: Cambridge University Press.

Canale, M (1983) On Some Dimensions of Language Proficiency, in Oller, J W (Ed.) *Issues in language testing research*, Rowley, MA: Newbury House, 333–42.

Canale, M, and Swain, M (1980) Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics* 1, 1–47.

Capel, A (2009) *A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists projects*, paper presented at the English Profile Seminar, Cambridge, 5–6 February 2009.

Chalhoub-Deville, M (2001) Task-based assessments: Characteristics and validity evidence, in Bygate, M, Skehan, P and Swain, M (Eds) *Researching pedagogic tasks*, London: Longman, 210–28.

Cizek, G J, and Bunch, M (2007) *Standard setting: A practitioner's guide*, Newbury Park, CA: Sage.

Clapham, C (1996) *The development of IELTS: A study of the effect of the Background knowledge on reading comprehension*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.

Coste, D (2007) Contextualising uses of the Common European Framework of Reference for Languages, Paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg 2007, available online www.coe.int/T/DG4/Linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc

Council of Europe (1998) *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference*, Strasbourg: Language Policy Division.

— (2001a) *Common European Framework of Reference for Languages: learning, teaching, assessment*, Cambridge: Cambridge University Press.

— (2001b) *European Language Portfolio (ELP)*, Strasbourg: Language Policy Division, available online www.coe.int/portfolio

— (2001c) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment – Guide for Users*, Trim, J (Ed.), Strasbourg: Language Policy Division, available online www.coe.int/T/DG4/Portfolio/documents/Guide-for-Users-April02.doc

— (2002) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Language examining and test development*, Milanovic, M (Dir.), Strasbourg: Language Policy Division, available online http://www.coe.int/T/DG4/Portfolio/documents/Guide%20October%202002%20revised%20version1.doc

— (2003a) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Manual, Preliminary Pilot Version*, Strasbourg: Language Policy Division, available online www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

— (2003b) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Manual, Overview of Preliminary Pilot Version*, Strasbourg: Language Policy Division, available online www.coe.int/T/DG4/Portfolio/documents/Overview.doc

— (2003c) *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Languages*, Strasbourg: Language Policy Division, available online www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html

— (2004) *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language examinations to the Common European Framework of Reference*, Strasbourg: Language Policy Division, available online http://www.coe.int/T/DG4/Portfolio/documents/CEF%20reference%20supplement%20version%203.pdf

— (2008) *Explanatory Notes to Recommendation CM/Rec (2008)7 of the Committee of Ministers to member states concerning the use of the Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*, available online https://wcd.coe.int/

— (2009) *Relating Language Examinations to the Common European*

Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual, Strasbourg: Language Policy Division, available online www.coe.int/t/dg4/linguistic/ Manual%20Revision%20-%20proofread%20-%20FINAL.pdf

Douglas, D (2000) *Assessing Languages for Specific Purposes*, Cambridge: Cambridge University Press.

Dudley-Evans, T and St John, M J (1998) *Developments in English for Specific Purposes*, Cambridge: Cambridge University Press.

Figueras, N and Noijons, J (Eds) (2009) *Linking to the CEFR levels: Research perspectives*, Arnhem: Cito/EALTA.

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, 261–279.

Figueras, N, North, B (Dir), Takala, S, van Avermaet, P, Verhelst, N (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual*, Strasbourg: Council of Europe.

Filipovič, L (2009) *English Profile – Interim report*, internal Cambridge ESOL report, April 2009.

Fulcher, G (1996) Testing tasks: Issues in task design and the group oral, *Language Testing*, 13/2, 23–51.

— (2004) Deluded by Artifices? The Common European Framework and Harmonization, *Language Assessment Quarterly* 1/4, 253–66.

Galaczi, E and ffrench, A (2007) Developing revised assessment scales for Main Suite and BEC Speaking tests, *Research Notes* 30, 28–31, Cambridge: Cambridge ESOL.

Galaczi, E and Khalifa, H (2009) Cambridge ESOL's CEFR DVD of speaking performances: What's the story? *Research Notes* 37, 23–9, Cambridge: Cambridge ESOL.

Grabe, W and Kaplan, R B (1996) *Theory and Practice of Writing: An Applied Linguistic Perspective*, Longman: London.

Green, T (2008) English profile: Functional progression in materials for ELT, *Research Notes* 33, 19–25, Cambridge: Cambridge ESOL.

Gunnarsson, B-L, Linell, P and Nordberg, B (Eds) (1997) *The construction of professional Discourse*, London: Longman.

Hayes, J R (1996) A New Framework for Understanding Cognition and Affect in Writing, in Levy, C M and Randsdell, S (Eds) *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, Mahwah, NJ: Lawrence Erlbaum Associates, 1–28.

Hawkins, J A (in preparation) Criterial Features of English across the Reference Levels of the CEFR, unpublished manuscript.

Hawkins, J A and Buttery, P (2009) *Criterial features in learner corpora: Theory and illustrations*, paper presented at the English Profile Seminar, Cambridge, 5–6 February 2009.

Hendriks, H (2008) Presenting the English Profile Programme: In search of criterial features, *Research Notes* 33, 7–10, Cambridge: Cambridge ESOL.

Horner, D and Strutt, P (2004) Analysing domain-specific lexical categories: evidence from the BEC written corpus, *Research Notes* 15, 6–8, Cambridge: Cambridge ESOL.

Ingham, K (2008) The Cambridge ESOL approach to item writer training: the case of ICFE Listening, *Research Notes* 32, 5–9, Cambridge: Cambridge ESOL.

Jones (2000) Background to the validation of the ALTE Can Do Project and the revised Common European Framework, *Research Notes* 2, 11–13, Cambridge: Cambridge ESOL.

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

— (2002) Relating the ALTE framework to the Common European Framework of Reference, in Alderson, J C (Ed.),167–83.

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

— (2009a) A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, in Figueras, N and Noijons, J (Eds), 35–43.

— (2009b) A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, *Research Notes* 37, 6–9, Cambridge: Cambridge ESOL.

Jones, N and Saville, N (2009) European Language Policy: Assessment, Learning and the CEFR, *Annual Review of Applied Linguistics*, 29, 51–63.

Kaftandjieva, F (2004) *Standard setting. Section B, Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment*, Strasbourg: Council of Europe, available online http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Khalifa, H and ffrench, A (2009) Aligning Cambridge ESOL examinations to the CEFR: issues and practice, *Research Notes* 37, 10–14, Cambridge: Cambridge ESOL.

Khalifa, H and Weir, C (2009) *Examining Reading: Research and practice in assessing second language reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Kurtes, S and Saville, N (2008) The English Profile Programme – An overview, *Research Notes* 33, 2–4, Cambridge: Cambridge ESOL.

Laming, D (2004) *Human judgment: The eye of the beholder*, London: Thomson.

Levelt, W J M (1989) *Speaking: from intention to articulation*, Cambridge, MA: MIT Press.

Linacre, J M (2006) Rasch Analysis of Rank-Ordered Data, *Journal of Applied Measurement*, 7/11, 129–139.

Little, D (2007) The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy, *The Modern Language Journal* 91/4, 645–55.

Louhiala-Salminen, L (2002) The fly's perspective: discourse in the daily routine of a business manager, *English for Specific Purposes* 21, 211–231.

Maris, G (2009) *Standard Setting from a Psychometric Point of View,* in Figueras, N and Noijons, J (Eds), 59–65.

Martyniuk, W (2008) Relating language examinations to the Council of Europe's Common European Framework of Reference for Languages, in Taylor, L and Weir, C J (Eds), 9–20.

— (Ed) (forthcoming) *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press.

McCarthy, M and Saville, N (2009) *Profiling English in the real world: what learners and teachers can tell us about what they know*, paper presented at American Association for Applied Linguistics Conference, Denver, Colorado, March 21–24 2009.

McNamara, T F and Roever, C (2006) *Language Testing: The Social Dimension*, Oxford: Blackwell.

Milanovic, M (2009) Cambridge ESOL and the CEFR, *Research Notes* 37, 2–5, Cambridge: Cambridge ESOL.

Morrow, K (Ed.) (2004) *Insights from the Common European Framework*, Oxford: Oxford University Press.

Mulholland, J (1999) E-mail: Uses, issues and problems in an institutional setting, in Bargiela-Chiappini, F and Nickerson, C (Eds), 57–84.

Nickerson C (1999) The Use of English in electronic mail in a multinational corporation, in Bargiela-Chiappini, F and Nickerson, C (Eds), 35–56.

North, B (Ed) (1992) *Transparency and Coherence in Language Learning in Europe: Objectives, Assessment and Certification*, Symposium held in Rüschlikon, Switzerland, 10–16 November 1991, Strasbourg: Council for Cultural Cooperation.

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

— (2002) A CEF-based self assessment tool for university entrance, in Alderson, J C (Ed.), 146–66.

— (2004) Relating assessments, examinations, and courses to the CEF, in Morrow, K (Ed.), 77–90.

— (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.

— (2007a) The CEFR Illustrative Descriptor Scales, *The Modern Language Journal* 91/4, 656–59.

— (2007b) *The CEFR Levels: Key points and key problems*, paper presented at the 23rd ALTE Conference, Sèvres, 18 April 2007.

— (2008) The CEFR levels and descriptor scales, in Taylor, L and Weir, C (Eds), 21–66.

North, B and Hughes, G (2003) *CEF Performance Samples. English (Swiss Adult Learners)*, available online www.coe.int/T/DG4/Portfolio/documents/videoperform.pdf

North, B and Jones, N (2009) *Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*, Strasbourg: Council of Europe.

North, B and Schneider, G (1998) Scaling descriptors for language proficiency scales, *Language Testing* 15, 217–62.

O'Sullivan, B (2006) *Issues in testing business English: The revision of the Cambridge Business English Certificates*, Studies in Language Testing volume 17, Cambridge: Cambridge ESOL/Cambridge University Press.

Papp, S (2007) *The Cambridge YLE tests in the light of cognitive, linguistic, educational, social-psychological and cultural aspects of children's L2 development and the CEFR*, internal Cambridge ESOL report.

— (2008) *Quantitative linking YLE to the CEFR – summary of empirical studies to date*, internal Cambridge ESOL report.

— (2009) Development of Can-do statements for KET and PET for Schools, *Research Notes* 36, 8–12, Cambridge: Cambridge ESOL.

Papp, S and Salamoura, A (2009) An exploratory study linking young learners examinations to the CEFR, *Research Notes* 37, 15–22, Cambridge: Cambridge ESOL.

Parodi, T (2008) *L2 morpho-syntax and learner strategies*, paper presented at the Cambridge Institute for Language Research Seminar, Cambridge, UK, 8 December 2008.

Rose, D (2008) Vocabulary use in the FCE Listening test, *Research Notes* 32, 9–16, Cambridge: Cambridge ESOL.

Salamoura, A (2008) Aligning English Profile research data to the CEFR, *Research Notes* 33, 5–7, Cambridge: Cambridge ESOL.

Salamoura, A and Saville, N (2009) Criterial features across the CEFR levels: Evidence from the English Profile Programme, *Research Notes* 37, 34–40, Cambridge: Cambridge ESOL.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C and Milanovic, M (Eds), 57–120.

— (2005) An interview with John Trim at 80, *Language Assessment Quarterly* 2/4, 263–88.

Shaw, S and Weir, C J (2007) *Examining Second Language Writing: Research and Practice*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Sherblom, J (1988) Direction, function and signature in electronic mail, *Journal of Business Communication* 25, 39–54.

Shohamy, E (2000) Assessment, in Celce-Murcia, M and Olshtain, E (Eds), *Discourse and Context in Language Teaching*, Cambridge: Cambridge University Press, 201–15.

Skehan, P (2001) Tasks and language performance assessment. in Bygate, M, Skehan, P and Swain, M (Eds) *Researching pedagogic tasks*, London: Longman, 167–85.

St John, M-J (1996) Business is Booming; Business English in the 1990s, *English For Specific Purposes* 15, 3–18.

Taylor, L (2003) The Cambridge approach to speaking assessment, *Research Notes* 13, 2–4, Cambridge: Cambridge ESOL.

— (2004) IELTS, Cambridge ESOL examinations and the Common European Framework, *Research Notes* 18, 2–3, Cambridge: Cambridge ESOL.

Taylor, L and Jones, N (2006) Cambridge ESOL exams and the Common European Framework of Reference (CEFR), *Research Notes* 24, 2–5, Cambridge: Cambridge ESOL.

Taylor, L and Weir, C (Eds) *Multilingualism and Assessment*, Studies in Language Testing Volume 27, Cambridge: UCLES/Cambridge University Press.

Thurstone, L L (1927) A law of comparative judgment, *Psychological Review* 3, 273–86.

Trim, J L M (Ed.) (1978) *Some possible lines of development of an overall structure for a European unit/credit scheme for foreign language learning by adults*, Strasbourg: Council of Europe.

— (Ed.) (1981) *Modern Languages (1971–1981)*, Strasbourg: Council for Cultural Cooperation.

— (Ed.) (2001a) *Common European Framework of Reference for Languages: learning, teaching and assessment – Guide for Users*, available online www.coe.int/T/DG4/Portfolio/documents/Guide-for-Users-April02.doc

— (2001b) *Breakthrough*, unpublished manuscript.

University of Cambridge ESOL Examinations (2006) *BEC Vantage Information for Candidates*, Cambridge: UCLES.

— (2007a) *First Certificate in English Handbook for teachers*, Cambridge: UCLES.

— (2007b) *Preliminary English Test Handbook for teachers*, Cambridge: UCLES

— (2008a) *Certificate in Advanced English Handbook for teachers*, Cambridge: UCLES.

— (2008b) *Certificate of Proficiency in English Handbook for teachers*, Cambridge: UCLES.

— (2008c) *Key English Test Handbook for teachers*, Cambridge: UCLES.

van Ek, J (1981) Specification of communicative objectives, in Trim, J L M (Ed.).

— (1987) *Objectives for foreign language learning: Vol. II Levels*, Strasbourg: Council of Europe.

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

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

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

Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

Weir, C J (2005a) Limitations of the Council of Europe's Framework of reference (CEFR) in developing comparable examinations and tests, *Language Testing* 22/3, 281–300.

— (2005b) *Language Testing and Validation: An Evidence-Based Approach*, Oxford: Palgrave.

Weir, C J and Milanovic, M (Eds) (2003) *Continuity and innovation: revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.

Wilkins, D A (1976) *Notional syllabuses*, Oxford: Oxford University Press.

— (1978) Proposal for Levels Definition, in Trim, J L M (Ed.), 71–8.

Williams, C (2007) *A preliminary study into the verbal subcategorisation frame: Usage in the CLC*, RCEAL, Cambridge University, UK, unpublished manuscript.

Wright, A (2008) A corpus-informed study of specificity in Financial English: the case of ICFE Reading, *Research Notes* 31, 16–21, Cambridge: Cambridge ESOL.

Framework articles in *Research Notes* – 2000 to 2009

Research Notes No	Title	Author(s)
Issue 2 (August 2000)	Background to the validation of the ALTE ‘Can Do’ project and the revised Common European Framework	Neil Jones
Issue 3 (November 2000)	BULATS:A case study comparing computer based and paper-and-pencil tests	Neil Jones
Issue 5 (July 2001)	Towards a common scale to describe L2 writing performance	Roger Hawkey
	The ALTE Can Do Project and the role of measurement in constructing a proficiency framework	Neil Jones
Issue 10 (November 2002)	Linking YLE levels into a single framework	Neil Jones
Issue 11 (February 2003)	Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts	Stuart Shaw
Issue 15 (February 2004)	Issues of test comparability	Lynda Taylor
Issue 16 (May 2004)	Exploring the relationship between YLE Starters and Movers and Breakthrough level	Trish Burrow
Issue 17 (August 2004)	Development of an Electronic European Language Portfolio	Simon Fenn
	Test Equivalence and Construct Compatibility across Languages	Peter Hardcastle
	A Common Solution to a Common European Challenge: The work of ALTE	Barbara Stevens
Issue 18 (November 2004)	IELTS, Cambridge ESOL examinations and the Common European Framework	Lynda Taylor
Issue 19 (February 2005)	The Common Scale for Writing Project: implications for the comparison of IELTS band scores and Main Suite exam levels	Roger Hawkey, Stuart D Shaw
	Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills	Neil Jones
Issue 21 (August 2005)	Listening, Reading and Writing on computer-based and paper-based versions of IELTS	Andrew Blackhurst
Issue 22 (November 2005)	Setting and monitoring professional standards: a QMS approach	Nick Saville
Issue 24 (May 2006)	Can Do self-assessment: investigating cross-language comparability in reading	Karen Ashton
	Cambridge ESOL exams and the Common European Framework of Reference (CEFR)	Lynda Taylor, Neil Jones
	Placing the International Legal English Certificate on the CEFR	David Thighe
	Linking learners to the CEFR for Asset Languages	Tamsin Walker
Issue 25 (August 2006)	Language testing for migration and citizenship	Nick Saville
Issue 27 (February 2007)	The comparability of computer-based and paper-based tests: goals, approaches, and a review of research	Neil Jones, Louise Maycock
Issue 33 (August 2008)	Building new corpora for English Profile	Theodora Alexopoulou
	Language pedagogy in an era of standards	Radmila Bodric
	English Profile: functional progression in materials for ELT	Tony Green
	Presenting the English Profile Programme: in search of criterial features	Henriette Hendriks
	The English Profile Programme – an overview	Svetlana Kurtes
	Directness, imposition and politeness in English and Russian	Tatiana Larina
	Aligning English Profile research data to the CEFR	Angeliki Salamoura
	Challenges to parsing English text: the language of non-native speakers	Caroline Williams
Issue 34 (November 2008)	The role of testing in an egalitarian society	Cecilie Carlsen
	Computer-based and paper-based writing assessment: a comparative text analysis	Lucy Chambers
	Views of Taiwanese students and teachers on English language testing	Jessica Wu
Issue 35 (March 2009)	A framework for migration and language assessment and the Skills for Life exams	Szilvia Papp, Martin Robinson
Issue 36 (May 2009)	Validating a worldwide placement test for German	Ardeshir Geranpayeh, Sibylle Bolton
	The classroom and the Common European Framework: towards a model for formative assessment	Neil Jones
	Development of Can Do Statements for KET and PET for Schools	Szilvia Papp
Issue 37 (Aug 2009)	Some evidence supporting the alignment of an LSP Writing test to the CEFR	Hugh Bateman
	Cambridge ESOL’s CEFR DVD of speaking performances: What’s the story?	Evelina Galaczi, Hanan Khalifa
	A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard-setting	Neil Jones
	Aligning Cambridge ESOL examinations to the CEFR: issues and practice	Hanan Khalifa, Angela ffrench
	Cambridge ESOL and the CEFR	Mike Milanovic
	An exploratory study linking young learners examinations to the CEFR	Szilvia Papp, Angeliki Salamoura
	Criterial features across the CEFR levels: evidence from the English Profile Programme	Angeliki Salamoura, Nick Saville

The above articles are searchable at: www.cambridgeesol.org/rs_notes/offprints/area/area_framework.html