English Profile Journal. (2010), **1**(1), Page 1 of 19, **e6** © Cambridge University Press 2010
doi:10.1017/S204153621000005X

Requirements for Reference Level Descriptions for English

Anthony Green Centre for Research in English Language Learning and Assessment, University of Bedfordshire

Abstract

The founding purpose of the English Profile Programme is to answer the Council of Europe's (2005) call for a set of Reference Level Descriptions (RLDs) for English linked to the Common European Framework of Reference for Languages (CEFR). The Council of Europe has issued guidelines setting out broad parameters for RLD development. This paper discusses how RLD might be developed for English in relation to the aims of the CEFR, incorporating consideration of critical voices, reports on the experiences of users of the CEFR and a review of currently operational RLDs for English: the Threshold *series. On the basis of these sources, recommendations are made for the ongoing development of the English Profile Programme.*

Keywords: RLD development, CEFR aims, learning objectives, proficiency scales

1. Introduction

The major goal of the Common European Framework of Reference for Languages (CEFR) (Council of Europe [CoE] 2001) is to provide 'a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks etc. across Europe' (CoE 2001: 1). Such has been the influence of the CEFR that 'across Europe' should perhaps now be extended to cover user groups from all parts of the world that wish to engage with the framework. To accomplish this goal, the framework aims to provide 'objective criteria for describing language proficiency [that] will facilitate the mutual recognition of qualifications gained in different learning contexts' (CoE 2001: 1).

However, this 'mutual recognition' is clearly not intended to imply a requirement for psychometric equivalence between the measures used to match learners to levels in these different contexts. There is no expectation that assessments used in different educational settings should be designed to a table of specifications directly derived from the CEFR, that they should employ similar item types in similar proportions, let alone that they should possess the equivalent score means, variance and reliability that such equivalence requires (see Feldt & Brennan 1989). Rather, the framework would seem to operate primarily through social moderation (Mislevy 1992); its coherent application is dependent on interaction and consensus building among users. If the users are unable to define and achieve agreement on the broad meaning of the levels, the framework loses its value as a basis for mutual recognition.

Despite criticism by Fulcher (2004); Krumm (2007); and Jones & Saville (2009) that some stakeholders use the framework as a means of imposing harmonisation – applying it to language programmes ‘as a hammer gets applied to a nail’ in Jones and Saville’s (2009: 54) phrase – this is not its intended purpose. Rather the framework is intended to allow for and to capture ‘the possible diversity of learning aims and the variety to be found in the provision of teaching’ (CoE 2001: 138). It is emphasised throughout the CEFR and its related publications that it is intended as a resource for consultation rather than a package for implementation and ‘should be open and flexible, so that it can be applied, with such adaptations as prove necessary, to particular situations’ (CoE 2001: 7).

The intended capacity of the framework both to describe multiple levels of ability and to cater to multiple contexts for language use is expressed by Richterich & Schneider (1992) through the concepts of *horizontality* and *verticality*. Horizontality is explained in the following terms: ‘[the] description and clarification of multi-dimensional content in terms of linguistic, social and cultural attainment, communication situations, or partial skills such as reading comprehension of texts of a certain type’ (Richterich & Schneider 1992: 44). Descriptors can be grouped horizontally in ways that are meaningful for specific audiences – learners, teachers, employers, government agencies – for use in diverse contexts. Learners placed at the same global level, but with different needs, may set themselves objectives based on – or be assessed against – descriptors drawn from quite separate CEFR tables.

A global scale, which provides holistic summaries of the levels, is presented in Table 4 of the CEFR in a form that might convey broad information to the ‘non-specialist user’ (CoE 2001: 24), while more detailed alternatives are suggested for learner self-assessment (CoE 2001: 26–27) and for the assessment of spoken performance (CoE 2001: 28–29). The self assessment scale presents distinctions between the skills of reading, listening, writing, spoken interaction and spoken production and the spoken performance scale further distinguishes within spoken language between such ‘qualitative aspects of language use’ as ‘accuracy’, ‘fluency’ and ‘coherence’ (CoE 2001: 25).

The system is also intended to be flexible with regard to the ‘vertical’ distinctions made between levels: the ‘quantity and quality of . . . skills’ (Richterich & Schneider 1992: 44). The six ‘reference levels’ in the CEFR are said to represent a ‘wide but by no means universal consensus on the number and nature of levels appropriate to the organisation of language learning’ (CoE 2001: 22). In the ‘branching approach’ suggested, a broader distinction can be made between three superordinate levels of learner (A: basic, B: independent and C: proficient) and finer distinctions can be made within the CEFR levels so that relatively small gains in language proficiency made within language programmes can be captured and reported. Within the scales presented in the CEFR, lines are sometimes drawn between descriptors representing the criterion level and those said to be ‘significantly higher’, although not sufficiently high to meet the demands of ‘the following level’ (CoE 2001: 36). This branching gives rise to levels such as B1.1 and B1.2 or B1 and B1+ and allows for further subdivisions so that ‘a common set of levels . . . can be “cut” into practical local levels at different points by different users to suit local needs and yet still relate back to a common system’ (CoE 2001: 32). The range of categories is said to allow for the construction of scales that reflect specific contexts for language use within the four specified 'domains': 'personal', 'public', 'occupational' and 'educational' (CoE 2001: 45).

The framework is said to be capable both of informing mastery decisions and of locating a performance on a continuum of proficiency (CoE 2001: 184). In other words, the descriptors are intended to be used both as a basis for specifying tasks that a learner at that CEFR level might be expected to work towards or succeed in performing – a 'constructor-oriented' purpose in Alderson's (1991) terms – and in providing differentiated descriptions of the quality of linguistic performance that can be used as rating scales to assign learners (all of whom might perform the same task) to the most appropriate level – an 'assessor-oriented' purpose (Alderson 1991).

2. Reference Level Descriptions

The Council of Europe (2005) recognises in the development of Reference Level Descriptions (RLDs) the need for a concerted effort to reconcile the competing aims of comparability and flexibility and to interpret and elaborate the necessarily broad descriptions of levels in the CEFR so that they may more readily inform practices relating to specific languages and to specific applications (such as syllabus design or test development). The essential requirements are set out in a set of draft guidelines as follows:

> for a given language, to describe or transpose the Framework descriptors that characterise the competences of users/learners (at a given level) in terms of linguistic material specific to that language and considered necessary for the implementation of those competences. This specification will always be an interpretation of the CEFR descriptors, combined with the corresponding linguistic material (making it possible to effect acts of discourse, general notions, specific notions, etc.) (CoE 2005: 4)

In common with other elements of what has come to be known as the CEFR toolkit, RLDs are intended to assist users in applying the CEFR to meet their local needs as language learners and educators.

The guidelines specify that RLDs should provide 'inventories of the linguistic realisations of general notions, acts of discourse and specific notions/lexical elements and morpho-syntactic elements' (CoE 2005: 5) that ground the CEFR descriptors. There is no requirement to limit the descriptors employed to those used in the CEFR; indeed, developers are encouraged to incorporate descriptors from European Language Portfolio (ELP) models. Each RLD should explain the process by which these inventories are arrived at, the knowledge of a linguistic form expected of learners at a given level (receptive or productive) and the relationships between the lists presented.

RLDs are thus intended to mediate between the CEFR and specific contexts for its use. They offer meaningful illustrative learning objectives that more fully operationalise the CEFR descriptions for users by providing linguistic exponents. In considering how best to proceed in developing RLDs, our first step should be to consider how learning objectives are presented in the CEFR and the extent to which these objectives already meet the needs of CEFR stakeholders as revealed through the available literature.

3. Objectives and 'can do' statements in the Common European Framework

The CEFR embodies an instrumentalist 'action-oriented approach'; the key concern is with what learners are able to accomplish when using a language rather than with their knowledge about language. This approach leads to an emphasis on potential language activities (and tasks) – 'any purposeful actions considered by an individual as necessary in order to achieve a given result in the context of a problem to be solved, an obligation to fulfil or an objective to be achieved' (CoE 2001: 10) – as meaningful outcomes of learning.

To assist in the process of setting objectives for language learning, teaching and assessment, the CEFR presents the user with conceptual questions for consideration and a bank of 'illustrative' descriptors arranged into 54 scales, each relating either to an aspect of a 'competence' (e.g. scales for 'vocabulary range' and 'vocabulary control' [CoE 2001: 112] are aspects of 'lexical competence' [CoE 2001: 30]) or to a language 'activity' (e.g. for oral production, 'Sustained monologue: describing experience', 'Sustained monologue: putting a case (e.g. in debate)', 'Public announcements', 'Addressing audiences'). These illustrative descriptors are arranged into six proficiency levels. According to the CEFR, 'The intention of providing a concrete illustrative set of descriptors, together with criteria and methodologies for the further development of descriptors, is to help decision-makers design applications to suit their contexts' (CoE 2001: 36).

It is acknowledged in the opening chapter of the CEFR that 'In order to fulfil its functions... a Common European Framework must be comprehensive, transparent and coherent' (CoE 2001: 7). The CEFR offers the following definitions:

> **Comprehensive**: '[specifying] as full a range of language knowledge, skills and use as possible'
>
> **Transparent**: 'clearly formulated and explicit, available and readily comprehensible to users'
>
> **Coherent**: 'free from internal contradictions. With regard to educational systems, coherence requires that there is a harmonious relation among their components'

The descriptors provided in the framework take the form of positively worded 'can do' statements describing activities that the learner might carry out in the target language. They are said to draw on three main sources: '(a) the theoretical work of the [CEFR] authoring group, (b) the analysis of existing scales of proficiency and (c) the practical workshops with teachers' (CoE 2001: 30). It is claimed that the scaling of these descriptors based on the rating of performance samples using the Rasch model (Bond and Fox 2007) creates an empirically derived scale of language ability.

The following section locates the CEFR in two descriptive traditions that influenced its development: behavioural objectives and the proficiency movement. The CEFR is considered in relation to its own stated purposes and in relation to the goals of these two approaches to level description.

4. Behavioural objectives and the proficiency movement

Behavioural objectives

Wilkins (1976: 13) makes the point that objectives of the kind espoused in the CEFR project are 'behavioural (though not behaviourist)' in the sense that they, in line with the 'action-oriented approach', are based on 'the purposes for which people are learning language and the kinds of language performance that are necessary to meet those purposes'. The following paragraphs provide a brief outline of how behavioural objectives have been used in educational contexts and the implications of the approach for the CEFR.

The mastery learning movement (Bloom 1968; Block 1971) is based on Carroll's (1963) conception of learning aptitude as representing variation between learners in the rate at which they might learn, not the eventual knowledge or ability that they might be able to attain. With its emphases on individuating learning – reducing variation in achievement by increasing variation in teaching (Bloom 1968) – and assessing progress formatively against clearly defined behavioural objectives, the movement appears, via its inception in a Council of Europe resolution on 'permanent education' (CoE 1970), to have had a strong impact on the development of the CEFR. This is evident from the attention given in the framework to the role of the learner and to achieving a coherence between language learning contexts: socially organised learning within language programmes and individual learning pathways across programmes or outside formal education (Schärer 1992).

An important element in mastery learning is the use of objectives or intended outcomes which specify the observable behaviours that should result from successful learning, appealing to authentic, 'real-world' applications (Block 1971). Carroll (1971: 31) argues that 'it is most essential . . . to be able to state as exactly as possible what the learning task is, particularly its objectives, in testable form. That is, a teacher must be able to determine when a student has mastered a task to a satisfactory degree'. Given such measurability, student progress may be tracked against objectives and feedback provided on task performance, with corrective instruction – employing alternative methods – provided to learners who do not achieve the intended objectives (Bloom 1968). Mastery learning ideas on the need for flexibility in teaching approaches are echoed by van Ek (1981: 17) in setting out the need for objectives in the Council of Europe programme, arguing that these should be explicit, but flexible, and allow 'maximum scope for differences between individual learners'.

In his classic text on the preparation of behavioural objectives, Mager (1991: 21 – first published 1962) suggests that three elements should be specified if an objective is to support mastery learning: performance, condition and criterion. The performance is a statement of what a learner should be able to do to demonstrate competence (the 'can do' element); the condition states relevant constraints under which the learner will be expected to perform; and the criterion sets out the level of performance that will represent success. An example provided by Mager (1991: 63) is: 'Given a DC motor of ten horsepower or less that contains a single malfunction, and given a set of tools and references [the condition], be able to repair the motor [the performance]. The motor must be repaired within forty-five minutes and must operate to within 5 percent of factory specifications [the criterion]'. Similar requirements continue to inform the design of 'outcomes statements', 'competencies' and 'standards' in general education – see, for example, Spady's (1994) requirements for 'content', 'competence' and 'context'.

The need for clarity in defining conditions (context) and criteria (competence) as well as performance (content) was recognised by van Ek (1987), who writes:

> In dealing with levels it will be convenient . . . to distinguish between what learners can do and how well they can do it, and to describe what learners can do in terms of the tasks they can perform and the language content they have at their disposal in performing them. In our further discussion, then, we shall adopt the following scheme, which, in fact, underlies most current level descriptions:
>
> | What | | How well |
> |---|---|---|
> | Task | Content | Quality |
>
> Task, content, and quality – it should be recognised – are not discrete parameters. What a person can do implies an ability to handle language content as well as a particular manner of doing it (quality). Yet, it is convenient to distinguish the three aspects in level descriptions, tasks being particularly described with a focus on discourse competence and content in relation to linguistic and sociolinguistic competence. Quality is a feature of all the various aspects of communicative ability.

Hence for van Ek (1987: 131), 'Level-descriptions are not complete without indications as to how well the learner is supposed to be able to perform the tasks specified in them'. North (2004) draws a parallel between this *what/how well* distinction in the CEFR and the notion of content and performance standards in standards-based education (Mislevy 1995). He equates 'what' with the scale categories of the CEFR and 'how well' with the hierarchy of levels, and suggests that, on the basis of content analysis and standards-setting exercises, courses and tests can be profiled against these standards. Experience with the draft Manual for Relating Tests to the CEFR (CoE 2003) suggests that this process may not be straightforward (Martyniuk forthcoming).

The proficiency movement

In language learning contexts, the scales of language proficiency developed in collaboration by the American Council on the Teaching of Foreign Languages and the Educational Testing Service (ACTFL/ETS 1986), building on the work of the US Foreign Service Institute (FSI) (Herzog n.d.), gave rise to what has been termed the 'proficiency movement' (Kramsch 1986; Clark & Clifford 1988).

Although originating in instruments designed for the judgement of performance in oral interview tests, their use was extended, in much the way that the CEFR is intended: to inform language programmes, textbooks and assessments (Liskin-Gasparro 1984). The ten-point ACTFL scales, like other widely applied language-testing schemes, came to represent a shared understanding of levels across language programmes so that educators felt that they had a clear idea of what was meant by a 'Novice High' or 'Intermediate Mid' level regardless of context (Chalhoub-Deville 1997).

Schemes that derive, like the ACTFL Proficiency Guidelines, from the work of the US Foreign Service Institute present a profile of typical learner abilities that are said to characterise a level. The '0+ (Memorized Proficiency)' on the ILR scale for Reading, is defined thus:

> Can recognize all the letters in the printed version of an alphabetic system and high-frequency elements of a syllabary or a character system. Able to read some or all of the following: numbers, isolated words and phrases, personal and place names, street signs, office and shop designations. The above often interpreted inaccurately. Unable to read connected prose.

The holistic nature of these scales requires the rater to arrive at a global judgement of a learner's level of language ability on the basis of observed performance. In the ILR scheme, the learner must fulfil *all* of the stated criteria in order to be judged to be at the level, while in other schemes such as the Australian Second Language Proficiency Ratings (ASLPR) the learner is ascribed to the level that appears to best fit their abilities, i.e. the performance need not satisfy all of the descriptors at the relevant level. The flexibility of the CEFR with its open-ended bank of scales means that it can only be compatible with this best-fit approach to global levels. However, this flexibility also raises questions. How close does the fit have to be – how many descriptors should apply – to justify reporting to a broad audience that a learner should be placed at one level rather than another? Are certain scales more central to the definition of a level than others? RLDs for English will need to further explore the relationships between the CEFR scale categories and the contribution these relationships make to defining the levels.

In common with behavioural objectives, proficiency scales describe performance in concrete functional terms to inform judgements about success: both describe outcomes of learning in North's (2004) terms – what a learner can 'actually do in the language' (Ingram 1996: 2), although the balance between descriptions of content (what/task) and descriptions of quality (how well) may vary across schemes. It is therefore unsurprising that proficiency scales, in spite of the objections of many applied linguists (Lantolf & Frawley 1985; Savignon 1985; Kramsch 1986), have come to be used as language learning targets – to 'represent a graduated sequence of steps that can be used to structure a foreign language program' (Liskin-Gasparro 1984) – and that behavioural objectives have been used in tracing progress along an assumed proficiency continuum. Initiatives such as the Graded Objectives in Modern Languages (GOML) movement in the UK (Page 1992) and the Dutch National Action Programme on Foreign Languages (van Els 1992) – both influential in the initial development of the CEFR – sought ways to integrate the verticality of language scales with the more horizontal orientation of learning objectives (Richterich & Schneider 1992) to represent expected levels of attainment for pupils at different stages of education and criteria for judging success.

5. Use of behavioural objectives and proficiency scale descriptors in the CEFR

Both proficiency scales and learning objectives were incorporated into the development of the CEFR illustrative descriptor pool (North 2000). The ACTFL Proficiency Guidelines and related proficiency scales such as the Australian Second Language Proficiency Ratings (Ingram 1990), the Foreign Service Institute Absolute Proficiency Ratings (Wilds 1975) and the Interagency Language Roundtable Language Skill Level Descriptions (ILR 1985) were included, as were such objectives-based schemes as the English National Curriculum: Modern Languages (1991) and the Eurocentres Scale of Language Proficiency (1993). Borrowing from both behavioural objectives and proficiency scales in this way sits well with the aim of building shared understandings between learner and teacher assessment of ongoing learning and consolidated summative information derived from programme external tests that may be used for certification and accountability.

The illustrative scales in the CEFR are made up of individual statements taken from the sources listed above, among others (30 schemes altogether), giving an initial pool of 1,679 descriptors in total (North 2000). The descriptors were screened for repetition and edited to produce 'positively worded, "stand-alone" statements that could be independently calibrated' (North 2000: 184).

Editing and trialling were intended to produce descriptors that would be suitable both as learning objectives and for distinguishing levels (North 2000). Five requirements for adequate proficiency descriptors emerging from North's study are listed in Appendix A of the CEFR (CoE 2001: 205–207):

Positiveness: worded 'in terms of what the learner can do rather than in terms of what they can't do'

Definiteness: 'describ[ing] concrete tasks and/or concrete degrees of skill in performing tasks'

Clarity: 'transparent, not jargon-ridden'

Brevity: 'a descriptor which is longer than a two clause sentence cannot realistically be referred to during the assessment process'

Independence: 'describe a behaviour about which one can say "Yes, this person can do this"'

Note the similarities between positiveness and independence and Mager's (1991) performance, and between clarity and definiteness and Mager's condition and criterion in formulating behavioural objectives.

Employing the Rasch model, the statements were then calibrated on the basis of teachers' ratings of student performance – either ratings of their own students or of video samples of learner speech. It is claimed that

> Because the illustrative descriptors constitute independent, criterion statements which have been calibrated to the levels concerned, they can be used as a source to produce both a checklist for a particular level, as in some versions of the Language Portfolio, and rating scales or grids covering all relevant levels, as presented in Chapter 3, for self-assessment in Table 2 and for examiner assessment in Table 3 (CoE 2001: 189)

Hudson (2005: 218) questions the validity of the claim of empiricism, observing that 'whereas the descriptors were empirically scaled based on performance ratings, the particular descriptors were not subsequently cast as actual test prompts and then calibrated again to determine if they still scale hierarchically'. North (2000) acknowledges the further objection that the scales are empirical only to the extent that they calibrate teacher perceptions: they are not empirically derived from L2 learner data (Hulstijn 2007). Reference level descriptions will need to address the links between teacher perceptions as operationalised in the scales and observable learner performance.

As the CEFR draws on both behavioural objectives and proficiency scales, it has attracted much of the criticism directed at both approaches as well as doubts concerning the extent to which it is possible to reconcile diverse purposes within a single scheme.

The criticisms made of the ACTFL Proficiency Guidelines (Lantolf & Frawley 1985; Savignon 1985; Kramsch 1986) that they were not based on research into the nature of second language acquisition have been repeated with respect to the CEFR by Bausch, Christ and Königs (2002, cited in Morrow 2004) and by Hulstijn (2007). Although, unlike the ACTFL Proficiency Guidelines, the CEFR is based on a componential model of communicative language ability, Hulstijn (2007: 666) notes the continuing lack of 'empirical evidence that, in following the overall oral proficiency scale, all learners first attain the functional level of A1, then the level of A2, and so on, until they reach their individual plateau'. Conversely, of course, there is equally little evidence to suggest that learners do not all proceed in this way.

Echoing criticisms of the ACTFL Proficiency Guidelines made by Bachman and Savignon (1986), Hulstijn (2007) notes that the CEFR levels integrate, through the 'activities' and 'competences' scales, issues of quantity – the number of 'domains, functions, notions, situations, locations, topics, and roles' (de Jong 2004) that a learner is able to cope with – and issues of quality – the extent to which communication is effective and efficient (although, as noted above, the degree of integration is inconsistent). Hulstijn questions the evidential basis for supposing that a learner placed at a given level on the basis of ability to perform the requisite activities will necessarily possess equivalent quality in terms of the linguistic (grammar and vocabulary) scales at that same level (2007: 666). Rather, he suggests, there may be learners who are limited in terms of quantity, but who are able to perform their limited range of roles with high linguistic quality, while others may be able to operate in a wide range of contexts, but with only limited linguistic resources: the distinction between horizontality and verticality made by Richterich & Schneider (1992) is not sufficiently clear in the CEFR.

The tasks that learners are asked to perform have a substantial impact on the nature and quality of the language that they will be able to produce. Further work is needed to tease out the relationship between elements of linguistic quantity and quality. The complex relationship between the two may partly explain the feeling that the higher levels of the CEFR are inappropriate or unattainable for certain groups of learners – especially young learners (Little 2007: 651) as the activities envisaged at these levels 'lie beyond the cognitive and experiential range of children and the great majority of adolescents'.

An objection raised by Weir (2005a) is that the framework is insufficiently explicit to inform test or task specifications because it fails to include the contextual parameters that affect performance – Mager's (1991) 'condition' or constraints on performance and Weir's (2005b) context validity. It is argued in the CEFR that 'very detailed concrete decisions on the content of texts, exercises, activities, tests, etc.' (CoE 2001: 44) will be required, taking into account the impact of physical conditions (e.g. levels of ambient noise), social conditions (e.g. relative status of interlocutors) and time pressures (e.g. conversational interaction allowing participants little time to prepare their utterances) on student/test-taker performance. Although a list of external conditions is provided (CoE 2001: 46–47), these are not incorporated into or explicitly related to the 'can do' descriptors that define the levels. As Weir (2005a) has pointed out, this must seriously undermine the interpretability of the levels as it is unclear which conditions should apply when we judge what learners 'can do'.

Furthermore, Weir (2005a) and Alderson (2007) argue, the scales fail to take sufficient account of the cognitive and metacognitive processes engaged by learners as they perform tasks at each CEFR level. These are dealt with in the framework (CoE 2001: 90–93), but as Alderson (2004) observes, the CEFR provides only 'a taxonomy of behaviours rather than a

Table 1 *Reading Benchmark 5: Initial intermediate proficiency (Pawlikowska-Smith 2000: 89).*

WHAT THE PERSON CAN DO	*EXAMPLES OF TASKS AND TEXTS*	*PERFORMANCE INDICATORS*
I. Social interaction texts Obtain factual details and inferred meanings in moderately complex notes, e-mail messages and letters containing general opinions and assessments of situations, response to a complaint and expressions of sympathy.	*C, S, W Read authentic notes, e-mail messages and letters (personal and public) containing general opinions, assessments of current affairs, response to a complaint/conflict, or expression of sympathy. Identify correctly specific factual details/inferred meanings.*	*Identifies specific factual details and inferred meanings in text. Identifies purpose of text, context of the situation, reader-writer relationship. Identifies mood/attitude of writer and register of the text.*

theory of development in listening and reading activities'. In the scale for 'Overall Reading Comprehension', by way of illustration, the descriptor 'Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension' not only fails to define what degree of comprehension is meant by 'satisfactory', but provides no details on the operations required to arrive at this level of comprehension. It is unclear, for example, whether, following the options set out on page 68 of the framework, learners should be able to read such a text for gist, for specific information, for detailed understanding or for implications, or whether such differences in reading purpose might have no implications for the level of the reading task.

The approach taken by the CEFR is contrasted by McNamara & Roever (2006) with schemes such as the Australian Certificates in Spoken and Written English (CSWE) or the related Canadian Language Benchmarks (CLB 2000) which provide elaboration and sample tasks to exemplify levels. Such elaboration is intended to help test developers, teachers responsible for assessment and learners themselves to arrive at comparable understandings of learner abilities. In the Canadian Language Benchmarks 2000 document, for example, the following 'can do' statement is given for the skill of reading at 'Benchmark 5: Initial intermediate proficiency'.

Note that Table 1 includes criteria for success as well as task definition and exemplification. A list of performance conditions is also provided – for Reading Benchmark 5 these include the following:

- Text is two or three paragraphs long and related to personal experience or familiar context.
- Text is legible, easy to read; is in print or neat handwriting.
- Tasks are in a standard format: with items to circle, match, fill in a blank, complete a chart, answer questions, etc.
- Learner is adequately briefed for focused reading (has at least minimal knowledge to activate knowledge schemata for top-down processing).
- Instructions are clear and explicit, for everyday situations, used with some visual clues, presented step by step. Pictures occasionally accompany text.
- Text has clear organization.
- Text is two or three paragraphs long, printed or electronic.

- Language is mostly concrete and literal, with some abstract words.
- Context and topic are often familiar and partly predictable for learner.
- Content is relevant and can be related to personal experience.
- Text types: newspaper articles, educational/content materials, stories, encyclopedia entries, short reports. (Pawlikowska-Smith 2002: 88)

In a companion volume (CLB Sample Tasks) examples of representative tasks are offered for each of the three CLB domains of 'community access', 'study/academic' and 'workplace'. The following reading tasks are suggested for the academic domain at Benchmark 5:

Study/Academic tasks

- Read a brochure about a training program.
- Follow instructions regarding school assignments.
- Follow simple-language, user-friendly computer screen commands.
- Number a set of pictures in an appropriate sequence based on the information in the text. Identify pictures that do not belong in a story sequence.
- Skim ten texts for 30 seconds each. Fill out a sheet with information on what each article is about, its purpose, if it is interesting or useful for you to read or not.

Guidance is also given on the proportion of items based on such tasks that a learner would need to answer correctly in order to achieve the benchmark.

The extent to which the CLB tasks might, in fact, yield comparable tasks and comparable results across teaching and learning contexts is, of course, open to question – how long after all are 'two to three paragraphs' and how might one be sure that a learner is 'adequately briefed'? There is insufficient information on required item types to ensure the comparability of tasks produced by different teachers based on these lists and so any assumption that 80% correct responses on one task would be even very roughly comparable to 80% on another seems overly optimistic. Nonetheless, the amplification of the 'can do' is far closer to the level of detail that Weir (2005a) has called for. Providing a similar level of detail for purposes of illustration might benefit the interpretability of RLDs for English.

The CSWE scheme goes even further than the CLB in supporting comparability in teacher assessment, providing sample tasks whose difficulty levels have been calibrated (using the Rasch model) for teachers to draw on when they assess learners (Brindley 2001). Of course, such calibration may be more meaningful when a scheme is designed to work within an educational programme (the case for the CSWE) than when it is intended to apply across programmes (the case for the CEFR and the proposed RLD).

Martyniuk & Noijons (2007) in their survey of the use of the CEFR across Europe found that users 'stress the need for general clarification (such as comments on theoretical concepts, examples and good illustrations, sets of tasks for use in specific contexts'. Further development of the sample materials provided by the Council of Europe or of open-access schemes such as DIALANG might help to ground the CEFR levels in a way that meets these needs. On the other hand, the more detail is provided, the greater the risk that *illustrative* tasks become *required* tasks and this distinction must be made clear.

The lack of guidance on criteria and conditions in the CEFR leads directly to questions about comparability. In posing the question, 'how does one know for certain that a test of Greek calibrated at level B1 in Finland is equivalent to a test of Polish considered to be at level B1 in Portugal?', Bonnet (2007: 670) points to two threats to comparability: local norms and inter-linguistic variation. In the absence of elaboration, exemplification and, most crucially, moderation of standards, it is likely that users in one setting may interpret the illustrative descriptors differently to users in another. It is possible that different interpretations of levels might develop so that, for example, a learner judged to be at level B1 in one school in Finland might be rated as level C1 at another in Portugal. The risks associated with such inconsistency are well illustrated by Crossey's (2009) account of the NATO STANAG 6001 scheme intended to provide agreed international language standards for military personnel. Faith in the scheme was undermined when it became clear that learners certified as being at a given level in one context did not satisfy the criteria as interpreted in another. Equally (although beyond the scope of a reference level description for English), the effect of language differences on the calibration of 'can do' statements is far from clear. It is important that RLDs incorporate a mechanism for ongoing interaction between users so that common understandings can be fostered and maintained.

Alderson is critical of inconsistencies in the wording of the descriptors, which perhaps reflect their varied provenance (Alderson *et al.* 2006; Alderson 2007). For example, at level C1 on the 'Coherence and Cohesion' scale (CoE 2001: 125) there is reference to 'clear, smoothly flowing, well-structured speech', but at level C2, 'coherent and cohesive text'. At level C1, learners use 'organisational patterns, connectors and cohesive devices'; level C2 only includes 'organisational patterns and cohesive devices'. However, in the global scale on page 28, 'connectors' do appear at level C2. On the scale for 'Vocabulary Range' (CoE 2001: 112), learners at level C have a 'lexical repertoire'; those at level B2 have a 'range of vocabulary'. Alderson *et al.* (2006) note that eight different verbs are used to indicate understanding at level B2, but that no gloss is provided to explain the implications, if any, of these differences. Without elaboration and exemplification it is unclear whether such differences are significant. RLDs for English should serve to clarify or minimise such variation.

The CEFR includes criticism of other scales for the use of vague language and qualifiers to define level differences, but such problems are not consistently avoided in the illustrative descriptors. In the 'Grammatical Accuracy' scale (CoE 2001: 114), a B2 learner 'Shows a relatively high degree of grammatical control', while a B2+ learner has 'Good grammatical control' and a C1 learner 'Consistently maintains a high degree of grammatical accuracy'. In attempting to apply these descriptions, the user may wonder, 'relative to what?' or 'how consistently?', but no guidance is offered in the CEFR. Descriptors for reading and listening make use of verbs such as 'understand', 'follow' or 'read'. These descriptors fall short on the criterion of clarity as they fail to specify observable behaviours that would demonstrate the learner's ability. Where comparisons are to be made across contexts it is important that users share similar expectations of learner performance.

There is considerable variation among descriptors even within scales in the extent to which they specify activities or refer to performance quality. In the self-assessment grid presented as Table 2 on pages 26–27 of the framework a B1 descriptor for spoken interaction reads, 'I can enter unprepared into conversation on topics that are familiar, of personal interest

Table 2 *Performance monitoring, evaluation and the Benchmark achievement report in CLB 2000 (Pawlikowska-Smith 2000: 146).*

1	*Fewer than 50% of the items*	*Performance not successful relative to task requirements; learner responds correctly to fewer than 50% of the items (comprehension questions)*
2	*Fewer than 70% of the items*	*Performance marginally successful relative to task requirements; learner responds correctly to fewer than 70% of the items (comprehension questions)*
3	*70–80% of the items*	*Performance successful relative to task requirements; learner responds correctly to 70–80% of the items (comprehension questions)*
4	*More than 80% of the items*	*Performance very successful relative to task requirements; learner responds correctly to more than 80% of the items (comprehension questions)*

or pertinent to everyday life (e.g. family, hobbies, work, travel and current events)', while a C1 descriptor is 'I can express myself fluently and spontaneously without much obvious searching for expressions'. The B1 descriptor, in common with many of those offered for the lower levels, describes a language activity ('conversation' – although no definition of a conversation is offered) and operative conditions (the speaker is 'unprepared' and the topics are 'familiar, of personal interest or pertinent to everyday life'). However, no criteria are suggested for the qualities of the learner's contribution to the conversation (nor are any provided for the other descriptor placed at this level). The C1 descriptor, in contrast, is exclusively concerned with quality or 'how well' ('fluently and spontaneously without much obvious searching for expressions'), but offers no guidance on activities or tasks beyond the vague 'express myself'. There is no indication of the conditions or situational constraints under which the speaker might be asked to interact.

Many case studies, in using the framework, call for further elaboration and exemplification of the levels (Alderson 2002; Morrow 2004; Martyniuk & Noijons 2007; Figueras & Noijons 2009; Martyniuk forthcoming). This need has been addressed by the Council of Europe through the provision of materials exemplifying the framework levels and the manual for relating language examinations to the CEFR. However, these materials remain sparse and seem to raise as many questions as they answer. North (2000) has suggested that 'Nobody has the same level across the 54 scales. Everybody has a profile': that the scales apply to aspects of competence rather than to learners. In performance a learner may, for example, display certain competences at level C1, others at B2. It is therefore important to understand which scales do apply to the sample performances or tasks and which do not. To what extent can a given level (rather than a profile) characterise a learner, a task or a performance? Sample performances provided to accompany RLDs for English need to be accompanied by commentary explaining in some detail how these relate to the descriptors.

6. Tension between purposes

Brindley (1998) lists challenges encountered in the development of outcomes-based assessment. Some of these are equally applicable to proficiency scales and include issues such as comparability and consistency of judgements outlined above. Additionally, there is a danger, also seen in the mastery learning experience of the 1960s and 1970s, that objectives can come to dominate the classroom. In finding an appropriate level of specificity in setting objectives, steering a course between the Scylla of inexplicit generalisation and the Charybdis of atomisation has proved to be a persistent challenge for educators (Popham 1973; Spady 1994; Brindley 1998). Almost every learning activity could be specified and recorded in terms of an objective, but this is administratively overwhelming and it is far from clear how information might be aggregated to provide information that might be of use to stakeholders outside the immediate learning context.

Within language programmes, teachers often employ relatively inconsequential 'intermediate objectives' (Trim 1981) to motivate learners and to acknowledge success on specific learning tasks; they may be less concerned with how performance on a task generalises to performance beyond the classroom. In contrast, external audiences such as sponsors and employers are interested in what are sometimes called 'terminal objectives' (Trim 1981) or summative representations of learner proficiency – the 'end-results of the learning process' (Trim 1981). In the CEFR context, the question arises of how many of the illustrative descriptors from the finer-grained category scales would need to apply before a learner could be judged to be at a certain level on the global scale (Council of Europe 2001: 24).

Teasdale and Leung (2000: 167) raise 'the question of whether assessments can at the same time adequately serve formative/diagnostic purposes whilst adequately performing as valid (in the psychometric sense) measurement instruments within particular contexts'. Arkoudis and O'Loughlin (2004) and Burrows (1998) point to the variation that teachers may exhibit in their interpretation of frameworks and to the ways in which they may adapt these to local circumstances in divergent ways. Such customisation is welcomed in the CEFR in its role as a heuristic for the elaboration of language programmes, but is likely to work against the comparability of outcomes from programmes purportedly situated at the same level.

A further objection to outcomes-based schemes is that they 'obscure, deform and trivialize education' (Egan 1983, quoted in Lantolf & Frawley 1985). Objectives are said to be bureaucratic and coercive, serving to circumscribe learner roles and forcing teachers to work towards predetermined ends. In the field of ESL, Auerbach (1986) criticises outcomes-based approaches for casting ESL learners in menial work roles and for narrowing the educational choices open to teachers. McNamara & Roever (2006: 212–213) are concerned that the CEFR is achieving such dominance that it imposes its interpretation of language learning on previously diverse teaching and testing programmes: 'testing organizations wishing to operate in Europe . . . have had to align their assessments with levels of the CEFR, despite radically different constructs, as a pure political necessity. Funding for reform of school language syllabi . . . is tied to conformity to the CEFR in more than one European country'. Morrow (2004) stresses that the framework is explicitly not intended as 'a set of suggestions, recommendations or guidelines' and cites the case studies that he introduces to the diversity of ways in which it may be applied. Although critical of some of these uses, Fulcher & Davidson (2007) take a less pessimistic view than McNamara & Roever (2006), seeing the framework's role in raising awareness as a key strength. In line with the ethos of the CEFR, RLDs for English should be presented as a resource for teachers, course developers and other educators and should be developed in consultation and collaboration with users.

It is apparent that the CEFR levels have provided a convenient set of standards that have been enthusiastically adopted by government agencies in mandating targets for achievement or setting requirements for immigration programmes, often without any clear justification (Alderson 2007; Krumm 2007). Alderson is concerned that

> there are claims that school leavers must achieve B1 . . . that university degrees in languages must be at level C2, and that migrants wishing to become citizens of a given country must attain level A2 (in the case of the Netherlands) or B1/2 (in the case of Denmark), without any thought being given to whether these levels might be achievable or justified (Alderson 2007: 662)

7. Operational Reference Level Descriptions

The *Threshold* series (van Ek and Trim) provides the classical approach to CEFR specification covering A1 (*Breakthrough*), A2 (*Waystage*), B1 (*Threshold*) and B2 and beyond (*Vantage*). These documents specify linguistic competence for English in a way that complements the CEFR. Rather than starting from language forms, as in the CEFR scales for range and control, the *Threshold* series starts from a classification of communicative functions and of notions, presenting lexical and grammatical forms as their exponents. *Profile deutsch* (Glaboniat *et al.* 2002), a reference level description for German, makes use of information technology to allow users to approach the material from either direction. In the *Profile deutsch* interface, clicking on a function brings up grammatical and lexical exponents; clicking on a grammatical item brings up a list of functions that the item could be used to realise. RLDs for English should make use of similar technologies to allow users to access content in directions that are of most relevance to their requirements.

Alderson (2004: 1), reporting on the experience of applying the CEFR scales to the DIALANG diagnostic assessment project, identifies a key challenge for users: ‘it is not easy to determine what sort of written and spoken texts might be appropriate for each level, what topics might be more or less suitable at any given level, and what sort of operation – be that strategy, subskill or pragmatic inference – or sociolinguistic context might be applicable at which level’. In addition to supplying language-specific specifications of grammar and vocabulary, the integration of cognitive and contextual parameters into the CEFR levels should be a key contribution of RLDs for English. Work contributing to the Council of Europe (2009) manual for relating examinations to the Common European Framework, such as the task content analysis checklists (Alderson *et al* 2006; Association of Language Testers in Europe 2005), has provided a framework for approaching such work. Shaw & Weir (2007) and Khalifa & Weir (2009) have taken up the challenge and have already established how contextual and cognitive parameters relate to the levels of Cambridge ESOL examinations. This work needs to be applied more broadly in RLDs to indicate how level differences are operationalised in a range of settings.

8. Conclusions

Martyniuk & Noijons (2007) concluded from their survey that there was ‘a considerable and quite urgent need to develop user-friendly sets of materials for mediating the

CEFR to the different stakeholder groups: policy makers, curriculum developers, textbook developers, publishers, teachers, testers, parents of learners, employers. There is also a strongly felt need for national and international cooperation in interpreting and using the CEFR'. Further RLDs for English will clearly need to extend the elaboration and exemplification of level-specific features of learner language, but also to encourage opportunities for exchange between users so that the social moderation of standards can be maintained.

To become more amenable to measurement – a requirement for educational as well as assessment purposes – the descriptors will need to be related to contextual and cognitive parameters. Extended specification of functions, notions, grammar and vocabulary with exemplars of the kind envisaged in the draft guidelines and exemplified by *Threshold* will not be sufficient to achieve this. In addition, there may a requirement for:

- extensive elaboration and perhaps revision of the terms used in the CEFR illustrative descriptors
- indications of the relationship between learner language skills and contextual parameters
- indications of whether and how the skills engaged by reading and listening tasks relate to the CEFR levels
- illustrative calibrated tasks
- support and guidance for application of RLDs to language programmes
- development of an international community of users with opportunities to share experiences and understanding of the CEFR levels

This paper has outlined the scope of the work that might be required to produce satisfactory RLDs for English. This indicates the need for a long-term and broadly based project. Over the short term, it may only be possible to realise elements of RLDs, but the production and validation of partial specifications is probably a necessary step towards more adequate tools in the future.

The English Profile partners have already begun the work of revisiting C-level functions (see Hawkey & Green 2008), but will need to prioritise the work of grounding the functional descriptions in evidence from learner language in developing effective RLDs for English.

References

Alderson, J. C. (1991). Bands and scores. In Alderson J. C. & North B. (eds.) *Language testing in the 1990s*. London: Macmillan, 71–86.

Alderson, J. C. (2002). Common European Framework of Reference for Languages: learning, teaching, assessment: case studies. Strasbourg: Council of Europe.

Alderson, J. C. (2004). Waystage and Threshold. Or does the emperor have any clothes? Unpublished manuscript.

Alderson, J. C. (2007). The CEFR and the need for more research. *Modern Language Journal 91*(4), 659–663.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment. Reading and listening. Final report of the Dutch CEF Construct Project. Unpublished document.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & 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.

American Council on the Teaching of Foreign Languages (1986). *ACTFL Proficiency Guidelines*. Hastings-on-Hudson: ACTFL.

Arkoudis, S. & O'Loughlin, K. (2004) Outcomes anxiety: ESL teachers assessing newly arrived ESL learners, *Language Testing*, *21*, (3), 283–303.

Association of Language Testers in Europe (2005). CEFR Grid for the Analysis of Speaking Tasks (report), Version 2.0, prepared by ALTE members. Retrieved on 22 February 2009 from www.coe.int/T/DG4/Portfolio/documents/ALTECEFRSpeaking GridOUTput51.pdf.

Auerbach, E. R. (1986). Competency-based ESL: One step forward or two steps back? *TESOL Quarterly*, *20*(3), 411–429.

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

Bachman, L. F. & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *Modern Language Journal 70*(4): 380–90.

Block, J. H. (ed.) (1971). *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment (UCLA-CSIEP) 1*(2), 1–12.

Bonnet, G. (2007). The CEFR and education policies in Europe. *Modern Language Journal 91*(4): 669–672.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programs: a review of the issues. *Language Testing 15*(1), 45–85.

Brindley, G. (2001). Implementing outcomes-based assessment: some examples and emerging insights. *Language Testing 18*(4), 393–407.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Burrows, C. (1998). Searching for washback: An investigation of the impact on teachers of the implementation into the Adult Migrant English Program of the assessment of the Certificates in Spoken and Written English. Unpublished PhD thesis, Macquarie University, Sydney, Australia.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record 64*(3), 723–733.

Carroll, J. B. (1971). Problems of measurement related to the concept of learning for mastery. In Block (ed.), 29–46.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing 14*(1), 3–22.

Clark, J. & Clifford, R. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status, and needed research. *Studies in Second Language Acquisition 10* (2), 129–147.

Council of Europe (1970). *Recommendation 611: On permanent education in Europe*. Strasbourg: Parliamentary Assembly.

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

Council of Europe (2003). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF). A Manual, Preliminary Pilot Version*. Strasbourg: Language Policy Division.

Council of Europe (2005). *Draft Guide for the Production of RLD: Version 2*. Strasbourg: Language Policy Division.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF). A Manual*. Strasbourg: Language Policy Division.

Crossey, M. (2009). The role of micropolitics in multinational, high-stakes language assessment systems. In Alderson J. C. (ed.), *The politics of language education: Individuals and institutions*. Bristol: Multilingual Matters, 147–164.

Davidson, F. & Lynch, B. K. (2002). *Testcraft: a teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (ed.), *Educational measurement*. New York: Macmillan, 105–146.

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

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

Fulcher, G. & Davidson, F. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teacher 40*(3), 231–241.

Glaboniat, M., Müller, M., Schmitz, H., Rusch, P. & Wertenschlag, L. (2002). *Profile Deutsch*. Berlin: Langenscheidt

Hawkey, R. & Green, A. (2008). English Profile Project 1: Functional progression at the C level of the CEFR. Unpublished project report.

Herzog, M. (n.d.). An overview of the history of the ILR language proficiency skill level descriptions and scale. Retrieved 22 February 2009 from www.govtilr.org/Skills/index.htm.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics 25*, 205–227.

Huhta, A., Luoma, S. Oscarson, M., Sajavaara, K., Takala, S. & Teasdale, A. (2002). A diagnostic language assessment system for adult learners. In Alderson J. C., Common European Framework of Reference for Languages: learning, teaching, assessment: case studies. . Strasbourg: Council of Europe.130–146.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal 91*(4), 663–667.

Ingram, D. E. (1990). Australian Second Language Proficiency Ratings. AILA Review 7: 46–61.

Ingram, D. E. (1996). The ASLPR: Its origins and current developments. Paper presented at the NLLIA Language Expo '96 (Brisbane, Queensland, Australia, July 19–21, 1996). Eric Document ED402735. Retrieved from www.eric.ed.gov 20 February 2009.

Interagency Language Roundtable (1985). *Interagency Language Roundtable Skill Level Descriptions*. Washington, DC: Government Printing Office.

Jones, N. (2002). Relating the ALTE framework to the Common European Framework of Reference. In Alderson (2002), 167–183.

Jones, N. & Saville, N. (2009). European language policy: Assessment, learning and the CEFR. *Annual Review of Applied Linguistics*, *29*, pp. 51–63.

de Jong, J. H. A. L. (2004). *Comparing the psycholinguistic and the communicative paradigm of language proficiency*. Presentation given at the international workshop 'Psycholinguistic and psychometric aspects of language assessment in the Common European Framework of Reference for Languages'. University of Amsterdam, 13–14 February, 2004.

Kaftandjieva, F. & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In Alderson (2002), 106–29.

Keddle, J. S. (2004). The CEF and the secondary school syllabus. In Morrow (2004), 43–54.

Khalifa, H. & Weir, C. (2009). *Examining second language reading*: *Studies in Language Testing*. Cambridge ESOL and Cambridge University Press.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal 70*(4) 366–72.

Krumm, H-J. (2007). Profiles instead of levels: The CEFR and its (ab)uses in the context of migration. *Modern Language Journal 91*(4), 667–669.

Lantolf, J. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal 69*(4), 337–345.

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals 17*(5), 475–489.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal 91*(4), 645–655.

Lowe, P. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and, particularly, to Bachman and Savignon. *Modern Language Journal 70*(4), 391–397.

Mager, R. F. (1991). *Preparing instructional objectives* (2nd edn.). London: Kogan Page.

Martyniuk, W. (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.* Cambridge ESOL and Cambridge University Press.

Martyniuk, W. & Noijons, J. (2007). Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States. Retrieved from www.coe.int 2 September 2009.

McNamara, T. F. & Roever, C. (2006). *Language testing: The social dimension.* Oxford: Blackwell.

Mislevy, R. (1992). *Linking educational assessments: Concepts, issues, methods and prospects*. Princeton NJ: Educational Testing Service.

Mislevy, R. (1995). Test theory and language-learning assessment. *Language Testing 12*(5), 341–369.

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

North, B. (ed.) (1992). *Transparency and coherence in language learning in Europe: objectives, assessment and certification*. Strasbourg: Council for Cultural Cooperation.

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

North, B. (2002). A CEF-based self assessment tool for university entrance. In Alderson (2002), 146–166.

North, B. (2004). Relating assessments, examinations, and courses to the CEF. In Morrow (2004), 77–90.

North, B. (2007). The CEFR illustrative descriptor scales, *Modern Language Journal 91*(4), 656–659(4).

Page, B. (1992). Graded objectives schemes. In North, B. (ed.), 64–67.

Pawlikowska-Smith, G. (2000). *Canadian Language Benchmarks 2000: English as a second language – for adults*. Toronto: Centre for Canadian Language Benchmarks.

Pawlikowska-Smith, G. (2002). *Canadian Language Benchmarks 2000: Additional sample task ideas*. Toronto: Centre for Canadian Language Benchmarks.

Popham, W. J. (1973). *Establishing Performance Standards.* Englewood Cliffs, NJ: Prentice-Hall.

Richterich, R. & Schneider, G. (1992). Transparency and coherence: why and for whom? In North (Ed.), 43–50.

Savignon, S. (1985). Evaluation of communicative competence: The ACTFL Provisional Proficiency Guidelines. *Modern Language Journal 69*(2), 129–134.

Schärer, R. (1992). A European language portfolio – a possible format. In North (ed.), 140–146.

Shaw, S. & Weir, C. J. (2007). *Examining writing in a second language*. Studies in Language Testing 26. Cambridge University Press and Cambridge ESOL.

Spady, W. (1994). *Outcomes-based education: Critical issues and answers*. American Association of School Administration: Arlington, Virginia.

Teasdale, A. and Leung, C. (2000).Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing*, *17*(2), 163–184.

Trim, J. L. M. (ed.) (1981). *Modern languages 1971–1981, Report presented by CDCC Project Group 4*. Strasbourg: Council for Cultural Cooperation.

van Ek, J. (1981). Specification of communicative objectives. In Trim, J. (ed.), Modern languages (1971–1981). Strasbourg: Council for Cultural Cooperation.

van Ek, J. (1987). *Objectives for foreign language learning: Vol. II – Levels*. Strasbourg: Council of Europe.

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

van Els, T. (1992). Revising the foreign languages examinations of upper secondary general education in the Netherlands. In North (ed.), 109–114.

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.

Weir, C. J. (2005b). *Language testing and validation: An evidence-based approach*. London: Palgrave Macmillan.

Wilds, C. P. (1975). The oral interview test. In R. L. Jones & B. Spolsky (eds.), *Testing Language Proficiency*, Washington, DC: Center for Applied Linguistics.

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