LANGUAGE ASSESSMENT QUARTERLY, *3*(1), 3–30

Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project

J. Charles Alderson
Lancaster University

Neus Figueras
Departament d'Educació, Catalonia

Henk Kuijper
CITO, The Netherlands

Guenter Nold
Deutsches Institut für Internationale Pädagogische Forschung, University of Dortmund

Sauli Takala
formerly of the University of Jyväskylä

Claire Tardieu
IUFM, Paris, Ministry of National Education, France

The Common European Framework of Reference (CEFR) is intended as a reference document for language education including assessment. This article describes a project that investigated whether the CEFR can help test developers construct reading and listening tests based on CEFR levels. If the CEFR scales together with the detailed description of language use contained in the CEFR are not sufficient to guide

Correspondence should be addressed to J. Charles Alderson, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YT, United Kingdom. E-mail: c.alderson@lancaster.ac.uk

test development at these various levels, then what is needed to develop such an instrument?

The project methodology involved gathering expert judgments on the usability of the CEFR for test construction, identifying what might be missing from the CEFR, developing a frame for analysis of tests and specifications, and examining a range of existing test specifications and guidelines to item writers and sample test tasks for different languages at the 6 levels of the CEFR. Outcomes included a critical review of the CEFR, a set of compilations of CEFR scales and of test specifications at the different CEFR levels, and a series of frameworks or classification systems, which led to a Web-mounted instrument known as the Dutch CEFR Grid.

Interanalyst agreement in using the Grid for analyzing test tasks was quite promising, but the Grids need to be improved by training and discussion before decisions on test task levels are made. The article concludes, however, that identifying separate CEFR levels is at least as much an empirical matter as it is a question of test content, either determined by test specifications or identified by any content classification system or grid.

This article describes a Dutch Ministry of Education, Culture and Science–funded project, accordingly dubbed the Dutch CEFR Construct Project. The purpose of the project was to develop an instrument, based on the Council of Europe's Common European Framework of Reference (CEFR; Council of Europe, 2001) as far as possible, which would describe the construct of reading and listening for English, French, and German which should underlie test items, tasks, and whole tests at the six main levels of the CEFR. (The CEFR contains three main levels—A, B, and C—each subdivided into two levels; thus A1 is the lowest level of proficiency described in the CEFR and C2 is the highest.) Such an instrument is intended to provide guidance to item writers on how to construct new test tasks and how to analyse existing test tasks at the various CEFR levels, as well as guidance to item bank builders on the design of item banks based on the CEFR levels and on how to select items for inclusion in such an item bank.

The CEFR is intended as a reference document for language curriculum and syllabus development, textbook writing, teacher training, and assessment. For details on the history and the development of the CEFR, see North (2000) and Council of Europe (2001, 2003). For accounts of case studies using the CEFR, see Alderson (2002) and Morrow (2004). The CEFR not only contains a comprehensive review of the elements that play a role in the teaching and learning of languages, but also includes numerous scales that describe a series of levels of language proficiency that have received considerable attention from professionals. The CEFR is increasingly referred to across Europe and claims are already being made that test X measures language ability at level Y on the CEFR. Therefore, an urgent need exists to illustrate the levels of the CEFR with calibrated test items. It is hoped that eventually it will prove possible to construct an item bank that can serve as a common operational tool that would enable the linking of national tests and examinations to the CEFR.

The experience of several previous projects, including the European Commission–funded DIALANG Project (Alderson, 2005; Alderson & Huhta, 2005), suggested that the CEFR in its current form may not provide sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level of the CEFR. Whereas the illustrative CEFR scales for the productive skills appear to be adequate for assessing written and especially spoken performance, in the case of the receptive skills the empirical evidence to justify the scales is not as strong and, therefore, the descriptors for receptive skills are unlikely to be sufficiently explicit and precise for test specifications to be developed. Thus the project team expected that further work would be necessary to make this possible. Such adaptation is envisaged and endorsed in the CEFR itself, as will appear from the following brief review of the CEFR approach.

To fulfil its functions, the CEFR was planned to be comprehensive, transparent, and coherent:

> By *comprehensive* is meant that the Common European Framework should attempt to specify as full a range of language knowledge, skills and use as possible … and that all users should be able to describe their objectives, etc., by reference to it.
>
> The CEFR should differentiate the various dimensions in which language proficiency is described, and provide a series of reference points (levels or steps) by which progress in learning can be calibrated.
>
> By *transparent* is meant that information must be clearly formulated and explicit, available and readily comprehensible to users.
>
> By *coherent* is meant that the description is free from internal contradictions. (Council of Europe, 2001, p. 7; page numbers refer to the 2001 English edition of the CEFR)

However, the CEFR emphasises that the construction of a comprehensive, transparent, and coherent framework for language learning and teaching does not imply the imposition of one single uniform system. "On the contrary, the framework should be open and flexible, so that it can be applied, with such adaptations as prove necessary, to particular situations (op. cit, p. 7)."

The Dutch CEFR Construct Project was carried out with this basic CEFR philosophy. It accepted the challenge of applying the CEFR to a special situation and making necessary adaptations. A key decision in the project was to exploit the CEFR as much as possible, identify gaps and areas that needed clarification, and produce a document that would serve the project goals. Project members acknowledged, however, that experience in issues other countries suggested, this would not be an easy task. In particular, attempts in the United States to use the American Council for the Teaching of Foreign Languages (ACTFL) guidelines for test development have resulted in controversy and criticism of the use of such frameworks to develop tests and test specifications (see Allen, Bernhardt, Berry, & Demel, 1988; and Lee & Musumeci, 1988, both cited in Alderson, 2000, pp. 278–281).

RESEARCH QUESTIONS AND TASKS

The CEFR, being a comprehensive description of language use, can also be considered, implicitly at least, as a theory of language development. However, the *can-do* scales for reading and listening present a taxonomy of behaviours rather than a theory of development in listening and reading abilities. Moreover, whether the still relatively abstract *can-do* descriptors in the CEFR can be turned into items that illustrate or exemplify the CEFR levels is far from clear. The experience of the CEFR-based DIALANG Project was that additional specifications needed to be developed before the CEFR could be used as the basis for test development. DIALANG is, however, only one example and is sui generis because it has developed diagnostic tests for delivery by computer across the Internet. It could not be assumed that the DIALANG experience and specifications would generalise across the variety of assessment contexts in Europe.

Thus, the basic questions the project asked were:

- Do we have in the CEFR an instrument to help us construct reading and listening items and tests based on the CEFR?
- If the CEFR scales together with the detailed description of language use contained in the document are not sufficient, what is needed to develop such an instrument, and what should the instrument be like?

After detailed inspection of the extent to which the CEFR itself serves as a basis for test specifications and whether it needs to be complemented and modified to eliminate ambiguities, the project planned to:

- develop a frame of analysis of items and tests of reading and listening;
- examine a range of items and tests claimed to be at the various CEFR levels;
- examine what the tests have in common in their test specifications and how they differ; and
- examine how the tests operationalise in test tasks the development of reading and listening abilities.

From such an investigation we hoped to develop a more specified theoretical framework and a practical instrument to complement the CEFR itself. The insights gained could lead to the development of guidance to test developers on how to analyse and construct items and tests of reading and listening at the various CEFR levels (A1 to C2).

The preceding research questions must be answered before attempts are made to link tests and examinations to the CEFR levels or before any potential European item bank for reading and listening is developed. Unless we have such an instrument, we will not, for example, be in a position to select suitable items for inclusion in an item bank on a principled (i.e., theoretical) basis rather than simply on psychometric criteria. What is needed is an instrument that contains test-relevant linguistic, psycholingusitic, and sociolinguistic as well as pragmatic criteria for text and task selection at different CEFR levels and for item construction or revision.

METHOD

Although solid theoretical foundations may be lacking, there is clearly a great deal of experience in producing tests and examinations at a range of different ability levels across Europe. Because many of these tests are explicitly claimed to be at various CEFR levels, it made sense to examine these tests and examinations to see what they had in common, how they differed, and how they operationalised in test items and tasks the development of language ability.

The project team, six language testing experts (the current authors) representing a range of testing and assessment cultures across Europe, convened to identify potentially relevant documents and to examine them for insights that could lead to the construction of a set of guidelines for test developers on how to construct both items and tests at the various CEFR levels. These experts have wide theoretical knowledge and practical experience in test construction, as well as being familiar with the diversity of assessment contexts in Europe and with using the CEFR in language education generally.

The method the project team used was iterative and inductive in the sense that the results obtained in each stage of the procedure were used to reflect on the outcomes, to plan the next stages, and to revise and extend the analytical tools as more experience was accumulated. The strategy in developing the analytical tools was first to adhere to the exact wording and listings of the CEFR and then to make adaptations as they were considered relevant. Continuous discussion, both focused and spontaneous, using e-mail and during meetings, was a crucial element of the project methodology.

The process began by a detailed inspection of what the CEFR had to say about reading and listening, compiling the CEFR scales for these skills by CEFR level (see Appendix A for a sample compilation for Reading [A1] and Listening [C1]), and then developing classification schemes based on these scales (see Appendix B for such a classification scheme for Reading at B2). These schemes were subjected to detailed analysis, criticism, and revision. They were then applied to sample test tasks whose CEFR level had previously been empirically established independently to ascertain to what extent the schemes were applicable. The first set of schemes was known as *frames*, but as the need became apparent for features or dimensions to be added to the classification schemes that were not explicitly contained in the CEFR, the name was changed to *Grid* to distinguish the latter from the entirely CEFR-based frames (of which Appendix B is one example).

These frames were revised twice and then adjusted and expanded into grids, which were themselves revised several times after analysis of rater agreement, taking into account the problems analysts reported and the suggestions they made for improvement. In addition, a selection of test specifications and guidelines to item writers for tests that had been empirically linked to the CEFR were also compared with the Grids to see to what extent the specifications described the development of reading and listening abilities in terms of the CEFR and the related Grids. The project team examined specimen tasks and specifications from DIALANG; the Dutch school-leaving examinations HAVO 2000 and MAVO 1999; the Profile Test Dutch as a Second Language; the Finnish Matriculation Examinations; the Finnish National Certificates for English; the Catalan Official Schools of Languages Examinations for English, French and German; the French Certificate of Higher Education in Foreign Languages; the Diploma of Language Competence; the Baccalaureate; Cambridge ESOL's Certificates in English Language Skills; Cambridge ESOL's Main Suite of English exams; and the Certificats de français and Zertifikat Deutsch produced by Weiterbildungs-Testsysteme GmbH (WBT).

RESULTS

The project resulted in several major outcomes: (a) an analysis and critique of the CEFR scales for reading and listening; (b) a Grid for the analysis of test items, texts, and tasks; (c) detailed information on the amount of agreement among individual analysts using later versions of the Grid; and (d) a compilation of the analysis of test specifications at the different CEFR levels using the Grid. As a result of these outcomes, a final version of the Grid was produced, and a brief users' guide and a more extensive training Grid with sample analyses were developed, together with an account of the usefulness of the Grid, and recommendations for its integration into empirical procedures for establishing CEFR levels of items and tasks. The Grid itself can be found at www.ling.lancs.ac.uk/cefgrid. (This current version of the Grid contains demonstration and training modules, as well as some dimensions that have been slightly revised or reworded to incorporate further analysis.) The following sections detail the project's main outcomes.

Analysis of the CEFR for Reading and Listening

The first outcome was a compilation of all CEFR reading and listening scales organised by level rather than by activity as they are currently presented in the CEFR, which facilitated a critique of the scales. (For reasons of space, the reader is referred again to Appendix A for a sample of such a compilation.)

The team extracted from the compiled scales and associated text features that appeared relevant to test design at any given CEFR level. For example, all the *can-do* statements begin with a verb that characterises aspects of the nature of comprehension (e.g., *understand*, *recognise*, *locate*, *infer*). Such features were termed *operations* and a category of *operations* constituted the first column in the frame of analysis intended to characterise what the CEFR says about comprehension at each level. Similarly, the *can-do* statements describe what somebody can comprehend at any given level, often in terms of the meaning of a text, the language of the text, and so on. The *source texts* that learners are said to be able to comprehend at any given level constitutes a third column in the instrument.

These compilations were used in developing and refining the frames and grids. These have incidentally proven extremely useful for familiarising analysts with the CEFR, and the compilations were also incorporated into the fourth and the final versions of the Grid. Most important, the process of producing the frames enabled the project members to focus on the wording of the CEFR descriptors and using the compilations repeatedly to construct and test the various frames and grids highlighted problems in the CEFR. The major problems were of four types:

1. Inconsistencies, where a feature might be mentioned at one level but not at another, where the same feature might occur at two different levels, or where at the same level a feature might be described differently in different scales.
2. Terminology problems: synonymy or not?
3. Lack of definition, where terms might be given, but are not defined.
4. Gaps, where a concept or feature needed for test specification or construct definition is simply missing.

These problems are illustrated in the following text.

Inconsistencies. Many formulations in the *can-do* statements are not consistent. Sometimes similar descriptions are found at different levels. Sometimes at one level (B1) something is said about vocabulary in texts, whereas at lower or higher levels nothing is said about vocabulary.

The operation *recognise* is only mentioned at levels A1, B1, and C1 and not at A2, B2, or C2. This cannot be a principled omission.

Despite the proliferation of verbs, inconsistencies are found in the use of different verbs nevertheless. *Infer*, for example, appears at some levels and not others, yet *inferencing* may well be needed even for A1 items. Certainly by B1 one would expect *infer* to appear as an operation, and therefore also at B2. Yet it only appears at C1.

The *use of a dictionary* is not mentioned in the CEFR at the lower levels—only at B2—yet lower levels are more likely to need to use a dictionary.

In *listening*, particular inconsistencies are found with type of speech:

- *Clear, slow, and carefully articulated speech* (A1).
- *Clear, slow, and articulated speech* (A2).
- *Clear, standard speech, familiar accent* (B1).
- *Normal speed, standard language* (B2).
- For C1 and C2 no limitations are set on speech.

Speed is not mentioned at B1; *standard* is first mentioned at level B1, but not at levels A1 and A2. A feature may appear in one descriptor for a level, but not in another for the same level. For example, what is the difference between *specific information* (A2) and *specific predictable information* (A2)?

We find a misleading inconsistency in the mention of specific text types at some levels in the CEFR but not at other levels. For example, advertisements are said to be processable at A2: *Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists, and timetables.* The only other reference to advertisements is at B2, where accommodation advertisements are specifically mentioned: *Can understand detailed information, for example, a wide range of culinary terms on a restaurant menu, and terms and abbreviations in accommodation advertisements.* But it cannot be the case that the ability to understand any advertisement (except for accommodation advertisements) is already developed at A2.

Similarly, *simple notices* can be read at A1 and *everyday notices* at A2. No other references to notices are found, but one can certainly envisage notices; for example, the regulations on permitted and forbidden activities in public parks, which are very hard to understand, albeit *everyday*.

Terminology problems: Synonymy or not? The CEFR uses a variety of verbs to indicate comprehension, some of which can stand alone and others of which require a noun phrase (see Table 1).

Often different words are being used synonymously, possibly for stylistic reasons or because the *can-do* statements were originally derived from a wide range of taxonomies. Thus:

TABLE 1
Verbs Used to Describe Comprehension

A2	*B1*	*B2*
Understand	Understand	Understand
Take	Locate	Scan
Get	Scan	Monitor
Follow	Identify	Obtain
Identify	Combine	Select
Infer	Extrapolate	Evaluate
	Recognise	Locate
		Identify

> *I can understand familiar names, words, and very simple sentences, for example, on notices and posters or in catalogues* (Self-assessment grid, Reading A1, CEFR, p. 26); and

> *I can recognise familiar names, words, and very basic phrases on simple notices in the most common everyday situations* (Reading for Orientation A1, CEFR, p. 70).

Clearly *understand* and *recognise* are synonymous here, but whether a meaningful distinction exists between the two main verbs in the following is unclear:

> *Can* identify *the main conclusions in clearly signalled argumentative texts,* and *Can* recognise *the line of argument in the treatment of the issue presented, though not necessarily in detail* (Information and argument, Reading for Information and Argument B1, CEFR, p. 70).

Are *find* and *locate* synonyms in the following:

> *Can* find *specific, predictable information in simple everyday material, such as advertisements, prospectuses, menus, reference lists, and timetables*, and

> *Can* locate *specific information in lists and isolate the information required, for example, use the Yellow Pages to find a service or tradesman* (Reading for Orientation A2, CEFR, p. 70)?

We decided to standardise the terminology and its consistent use as much as possible. Because the CEFR has no description of how cognitive operations might differ at different levels (or even whether they do), there is no basis in the CEFR itself has no basis for standardising or grouping verbs, and we had to have recourse to theories of comprehension to resolve this issue.

Lack of definitions. Many terms are used in the CEFR, but they are undefined. For example, *simple* is frequently used in the scales, but how one is to decide what is *simple* compared to what is *less simple* and, especially, what is *very simple* is not clear. The CEFR is language independent, and thus does not contain any guidance, even at a general level, of what might be simple in terms of structures, lexis, or any other linguistic level. Therefore, the CEFR would need to be supplemented with lists of grammatical structures and lexical items for each language to be tested, or it could recommend the use of electronic corpora, which could be referred to if terms such as *simple* and *frequent* are to have any meaning for item writers or item bank compilers. Of course, what is simple for one first language background might be far from simple for somebody with a different first language, and therefore some appeal must be made to second language acquisition (SLA) theory or research. This may prove to be an intractable problem for tests intended for multilingual audiences.

The same definitional problem applies to many expressions used in the CEFR scales: for example, *the most common*, *everyday*, *familiar*, *concrete*, *predictable*, *straightforward*, *factual*, *complex*, *short*, *long*, *specialised*, *highly colloquial*, and doubtless other expressions. These all need to be clarified, defined, and exemplified if items and tasks are to be assigned to specific CEFR levels.

However, what is familiar in one culture with particular background knowledge and expectations, may not be at all familiar in other cultures (or individuals). How this can be taken into account by item writers is far from clear, even though it may make sense in a self-assessment scale because individual respondents can decide for themselves what is familiar, everyday, or specialised. Even so, individuals cannot decide for themselves what is short or long.

Gaps. We considered a feature missing if it was mentioned in general terms somewhere in the CEFR text but then was not distinguished according to the six CEFR levels or was not even specified at one level.

The first major gap in the CEFR we identified, as noted previously, was a description of the operations that comprehension consists of and a theory of how comprehension develops.

Related to this is the absence of any specification of microskills or subskills of comprehension. The one most immediately noticed was *skim*, but others such as *distinguish relevant from irrelevant details* or *discriminate between fact and opinion* also seem to be absent.

The text of the CEFR introduces many concepts that are not then incorporated in the scales or related to the six levels in any way. These include the following: competence, general competence, and communicative language competence (pp. 9, 13, 108ff.); activities, processes, domain, strategy, and task (pp. 10, 14, 15, 16); context (pp. 48–49, Table 5); ludic and aesthetic uses of language (pp. 55–56); text-to-text activities (p. 100); sociocultural knowledge (pp. 102–103); study skills (pp. 107–108); tasks, including description, performance (conditions, competencies, linguistic factors), strategies, and difficulty (pp. 157–166).

One major element missing from the CEFR is the task: what do candidates have to do with text? Although an entire chapter of the CEFR is devoted to this topic, at no point is there a discussion of how tasks might be distinguished by level. In fact, some of the illustrative scales are indeed subdivided by task in a sense because they address things such as:

- listening as a member of a live audience;
- reading for orientation; and
- reading for information and argument.

But other illustrative scales address specific texts:

- listening to announcements and instructions;
- listening to radio and audio recordings; and
- reading instructions.

In short, we find no principled way in which such illustrative scales have been created, and the dimension of purpose—why one is reading or listening to any given text in any particular setting—is not addressed systematically at all. This gap is a serious problem for test writers and item bank builders.

In tests, however, in a sense the test method is the task, and so a consideration of test methods is crucial. For multiple-choice methods in particular, the nature of the options offered should be considered part of the text to be processed, but distinguished from the input text. How the options are constructed, what content they have, how they are worded, in what order they appear, how many pieces of information in a text they address: all these and more will necessarily add to the difficulty or ease of the item, but the CEFR currently has no way of taking this into account because it focuses exclusively on *action* and *real-world* use. Discussion of test method is absent from the CEFR. However, although item writers need to know what test method to use at which CEFR level, such method effects will most likely generalise to more than one CEFR level, and they are unlikely to be defining characteristics of listening or reading tasks at one level and not another. Nevertheless, the processing demands they create need to be taken into account somehow when devising specifications and giving guidance on what level a *performance* is at.

The Final Grid

Given that the frames were based entirely on the CEFR, the same problems identified in the CEFR were necessarily also contained in the frames. Because it was necessary to fill the gaps in the CEFR and the frames, a new instrument dubbed a

Grid was developed, which added dimensions that appeared to be needed, even though they do not appear in the CEFR.

Five cycles of reformulating initial grids as a consequence of discussions about the content and applying the successive grids to texts, tasks, items, and specifications resulted in the Final Grid (see Appendix C for the content of the Final Grid).

The Grid is itself divided into three parts: (a) the text(s) on which the test is based; (b) the items; and (c) the whole task, which consists of a combination of text(s) and items.

The following sections outline the dimensions of the Final Grid, which derive directly from the CEFR.

Text. The text consists of:

- text source;
- topic;
- domain; and
- CEFR level.

Eleven dimensions can be seen as a generalisation of descriptions that are found in the CEFR but with inconsistent formulations:

- authenticity (i.e., not abridged or simplified texts);
- discourse type;
- nature of content (the abstraction dimension);
- text length (words for reading, duration for listening);
- vocabulary;
- grammar;
- text speed (only for listening);
- number of participants (only for listening);
- accent/standard (only for listening);
- clarity of articulation (only for listening); and
- how often played (only for listening).

In addition, raters were asked to judge at which CEFR level a learner would have to be to find the text comprehensible.

Item.

- Operations: The CEFR lacks descriptions of operations as we have seen. Therefore, we had to develop a description of the operations that not only was consistent and related to theories of listening and reading, but also that still

related to the descriptions in the CEFR *can-do* statements as far as possible. This resulted in the three-part description of the mental operations:
  - the behaviour: recognising, making inferences, and evaluating;
  - the source of the information: whether it is explicit in the text or only implicit; and
  - the "what" is understood—a category describing what is to be read or listened for.
- Item types were taken from standard textbooks on language testing and assessment.
- Estimation of the CEFR level of the item: based on the compilations of CEFR scales by (a) level, (b) activity, (c) DIALANG can-do, and (d) ALTE levels.

Task. The *task* was a simple combination of text(s) and item(s) with an estimate of the CEFR level of the task as a whole. Whereas *item level estimated* applies to the estimated level of an individual item and its associated text(s), *task level estimated* is an expert judgment of the overall level (weighted or unweighted, depending on the individual rater's judgment) of all those items based on the text(s) and grouped into one task.

Agreement on Using the Grid to Analyse Texts, Items, and Tasks and to Estimate CEFR Levels for Each

The Final Grid was used both to characterise a range of tasks from different sources and also to serve as a framework for the analysis of the test specifications we received from various exam bodies. It was considered important for all analysts to complete all analyses as a further test of the transparency and applicability of the Grid, and therefore all items selected had to be in English. Accordingly, the project coordinator selected reading tasks from the following sources whose items had been empirically analysed and related to the CEFR:

- Cambridge ESOL: PET (= B1);
  FCE (= B2);
  CPE (= C2) [sample tasks in publicly available handbooks];
- Catalan Official Schools of Languages Exams:
  Elemental (= B1);
  Aptitud (= B2); and
- Finnish Matriculation Examinations: mixed levels.

Two tests were selected from each level available, anonymised, placed in random order, and then compiled into a booklet of 77 items, giving 16 tasks (Finnish, 6; Catalan, 4; Cambridge, 6). Each analyst was asked to complete the Grid without discussing results with colleagues.

Average agreement among individual analysts of more than 75% was achieved on the dimensions *authenticity*, *domain*, and *broad discourse type*, and *text source* came close at 73.75%. Agreement of less than 60% only occurred on *task level estimated.* This was notably better than had been achieved with earlier drafts of the Grid. The only dimension where agreement was less was on *topic* (draft 3, 62.5%; draft 2, 77.6%). However, striking differences exist among analysts in terms of the frequency of use of the grid dimensions, especially for *item type*, *text source*, *topic*, *vocabulary*, *grammar,* and *operation.* Clearly, different analysts used the categories differently. Individual input to a grid will most likely result in disagreement and discrepancy; therefore, it is essential that Grid users receive familiarisation and training in using the Grid, as well as examples of exponents of any dimension where possible. The provision of such (agreed) examples was, however, not possible until the final phase of the project. Nevertheless, and despite this level of disagreement, completion of the grid by groups of individuals could clearly facilitate useful comparisons of results and discussions of the reasons for the different perceptions. This in itself could lead to enhanced understanding of the CEFR and of the categories of the Grid.

Agreement among analysts on item CEFR levels was significant, ranging from a moderate .49 to a high of .78. Individual analysts' agreement with empirically established item and task levels was, however, only moderate—in the .50s and .60s. Such relatively modest correlations show again the need for training, team discussions, and team decisions when inputting data to the Grid.

Analyses were conducted separately for dimensions pertaining to items and those pertaining to texts. Chi-squares were calculated to test the strength of associations between dimensions and CEFR levels, but most did not meet the necessary levels of expected cell frequencies and results can be seen only as tentative. Nevertheless, the project team members considered that they enabled the development of initial hypotheses about relationships, which would have to be falsified in further research using a greater number of items and tasks.

Item type showed no association with CEFR levels, and the *operations* judged to be tested by items varied considerably across analysts, sometimes reaching significance, sometimes not. Analysts 3 and 5 agreed that the most common operation was *recognise and retrieve explicit detail,* whereas Analyst 2 found *recognise and retrieve explicit main idea/gist* to be the most common. Analyst 1 disagreed, but agreed with Analyst 4 that the most frequently occurring operation was *Evaluate implicit text structure/connections between text parts.* Overall only moderate agreement existed among analysts with regard to what individual items were testing. Substantial differences also occurred among analysts regarding which operations they identified in the various items. This is in keeping with findings in the literature on reading in a foreign language (see Alderson 2000), but it presents considerable difficulties for those who wish to claim that CEFR levels can be distinguished by operations or skills, and it underscores the finding that reaching agreement on what operations are required by any given item at any CEFR level is rather difficult.

However, when the CEFR levels were grouped into three broader levels (A, B, and C) the results showed a tendency for items at lower levels to be more focused on retrieving explicit information from texts, whereas at higher levels inferring from and evaluating texts became more prominent, and items tended to deal more with implicit information. Thus, some hope is provided by rather coarser-grained analyses at three broad CEFR levels, reinforcing the desirability of further research using larger samples of texts and items to explore possible relations, especially if the analysts are trained in advance and discuss their analyses among themselves before reaching final decisions.

With respect to text characteristics, no significant association was found between the CEFR level of a text/task and *authenticity, domain, grammar, text source, discourse type, topic,* or *degree of abstractness of content.* The only dimension that showed a significant association was *vocabulary.*

However, although this analysis was applied to 77 items belonging to 16 tasks, we must stress that more extensive research using the Grid is needed before solid conclusions can be reached about the relation, or lack thereof, between the dimensions of the Grid and CEFR levels. Our results can be considered only suggestive, given the limited time available for this project.

The Analysis of Test Specifications

In addition to analysing test tasks and items, project members analysed the test specifications to which they had access, using the dimensions of the latest draft of the Grid. The aim of this procedure was to see to what extent tests at the same level of the CEFR, produced by different examining bodies, agreed in content and specifications.

Test specifications from test providers from the United Kingdom, Catalonia, the Netherlands, Finland, Germany, and France were examined. These specifications related to numerous European languages. Although a detailed report of the results is beyond the scope of this article, results can be found in appendices 12 and 13 of Alderson et al. (2004), and the analysis led to some important conclusions.

First, the Grid was a useful instrument to describe and analyse the test specifications examined. It could therefore perhaps be used to analyse the diverse practices in language testing across Europe, and it offers a tool for describing relations between specifications and the CEFR. Second, in addition to general similarities, we found many differences in the way test specifications dealt with descriptions of the characteristics of input texts and items for listening and reading and in the terminology used. Third and importantly, however, there appear to be no systematic differences in the test specifications examined in terms of most of the dimensions included in the Grid as CEFR level changes. The specifications examined barely distinguish among CEFR levels in terms of content.

Indeed, the specifications analysed do not seem to be based on a theoretical construct—on how the language to be tested is understood. Some specifications appear to have been written focusing on the details of exam format and length for a particular level without seriously considering language proficiency as a whole or the development of reading and listening proficiency from beginning to advanced levels. This focus on exam format at the expense of theory or construct may well be the reason why there is a lack of systematic and clear use of terminology as well as a lack of uniformity of style and approach across levels.

Most important for this project, there is very little information on how different dimensions may affect difficulty or how the dimensions may vary across CEFR levels. A common understanding of the specifications by item writers seems to rely in most cases on exemplification (previous exams) and local expertise. This suggested the need—in addition to item writer training—to provide illustrative examples for the Grids to guarantee a common understanding of whatever terms or labels are used. These were provided in the final phase of the project.

IMPLICATIONS FOR USING THE GRID

From the iterative process of creating frames, critiquing and revising them, applying them to texts and tasks, and making further revisions, what was emerging was an instrument that was becoming increasingly user-friendly, which was still based on the CEFR, but which had added important dimensions that need to be included in any test specification. Additionally, as we saw in the previous section, many of the test specifications we analysed did not adequately relate to the CEFR or to theories of comprehension; above all they did not distinguish among their various targeted levels in any consistent way. Using the Grid clearly highlighted this conclusion, and in future the Grid could be very useful in helping test developers relate their test specifications more closely to the CEFR.

Furthermore, from the application of the Grid to texts and items, it became clear that the Grid could indeed be used to describe such items, texts, and tasks in terms of the CEFR, but with added dimensions. The best way of reaching agreement on the description of texts and items clearly was for analysts initially to attempt their individual analyses of the texts and items, but then to convene to discuss the individual analyses, identify sources of disagreement, and resolve differences before deciding on the definitive analysis of the texts and items.

As a result, the Grid will also be very useful in the various processes recommended in the Manual (see Council of Europe, 2003; and Figueras, North, Takala, Van Avermaet, & Verhelst, 2005) for linking the tests to the CEFR. In particular, the Grid could prove useful in the familiarisation, specification, and standard-setting stages of the Manual. Indeed, it has already proved useful in workshops introducing the CEFR to teachers and testers, and the associated training module has proved particularly effective in this regard.

In addition, although the project has been critical of aspects of the CEFR, the Council of Europe itself, as well as many European institutions and projects, have already begun to use the Grid when seeking to link their tests to the CEFR levels. In particular, the Council of Europe has developed a CD-ROM that is intended to exemplify CEFR levels of reading and listening in which the Grid is used to characterise the content of reference materials for reading and listening in English, French, German, Spanish, and Italian (Council of Europe, 2005). A European Commission-funded project (EBAFLS; Gille & Sluiter, 2005) to develop a European item bank for reading and listening has also incorporated a version of the Grid in its classification of test items and tasks.

Finally, the Grid clearly will prove very useful in conducting research into the question of which dimension or dimensions of the Grid (and thus of the CEFR) contribute most to the level of reading and listening texts and items once a sufficiently large database of analysed texts, items, and tasks has been assembled. This could lead to a growing clarification and understanding of what the CEFR levels for reading and listening are, what further dimensions might need to be added, and how these abilities develop. Given that the problems this project faced are similar to those experienced elsewhere (cf. the use of the ACTFL guidelines discussed in Alderson, 2000, p. 278–281), such a database would be of considerable interest internationally.

LIMITATIONS OF THE GRID: THE NEED FOR AN EMPIRICAL PROCESS

The basic conceptual problem the project team faced was determining what it meant for an item to be at a certain level. People are said to be at B1 if they can do the things described at that level to a satisfactory degree, but not (yet) most of the things they should be able to do at the next higher level. The problem this project was intended to address, however, was what the construct validity of concepts such as B1 is. A test developer has to show that he or she can build measuring instruments that can classify people at the level to which they truly belong. However, this presents a circular problem: to validate the theory, measurement instruments are needed, but to validate these measurement instruments, a theory is necessary on which one can rely.

Moreover, to know whether a given item is indeed at the level of difficulty intended, piloting on suitable samples of test-takers is crucial. But to do this, a suitable sample is needed (i.e., knowing the level of the test-takers is necessary to judge the adequacy of the item). A problem of circularity therefore presents itself.

To escape this circle, the project team proposed the following procedures:

- Describe the text and tasks using the dimensions of a classification system (the Grids).
- Make a guess at the level of a task (guided by the classification system and the CEFR scales), leading to an estimated CEFR level.
- Pretest the tasks thus labelled, describing in detail the characteristics of the pilot sample.
- Calibrate the tasks.
- Do standard setting to set the boundaries of the levels on the scale coming from the calibration.
- Assign a psychometric level to the tasks.
- Assign a definitive level to the tasks, which is possible only if the psychometric level falls within the band of the estimated level (in other words if the estimation based on the *analysed content* is comparable with the *psychometric* value).

In short, the identification of separate levels in the CEFR is at least as much an empirical matter as it is a question of the content of the tests as determined by test specifications or as identified by our Grids. However, the project team member suspect that examining the linguistic characteristics of texts and tasks in much more detail than has been possible in this project will be necessary if adequate characterisations are to be identified of the content and construct of tests of *reading* and *listening* at the different CEFR levels. This is likely to involve:

1. identifying tests and tasks that have been incontrovertibly scaled on the CEFR;
2. developing measures of the linguistic features of texts and tasks that previous research has shown to be relevant to defining difficulty, independently of the CEFR (see, e.g., Alderson, 2000; Buck, 2001; Buck et al., 1997; and Shiotsu & Weir, 2004); and
3. applying such measures experimentally to the texts and tasks identified in the first step to see to what extent analysis of the linguistic features of such texts and tasks can predict CEFR levels. This will most likely be an extensive project, and we therefore recommend that it be conducted first for *reading* for two languages only—English and French—and that, if successful, the research be later extended to *listening* for the same languages, and only then in a third phase to other languages.

CONCLUSIONS

The Dutch CEFR Construct Project developed a framework based on the CEFR for analysing language test items, texts, tasks, and specifications to help test developers relate their examinations to the CEFR. This framework has been turned into a Web-based Grid, which is completed by analysts and whose data goes into a database that facilitates the analysis of results from the point of view, inter alia, of the amount of agreement among analysts on the content of the test items, tasks, and so forth.

This project did not intend to conduct extensive research into what makes reading and listening tests difficult, but rather sought to develop, on the basis of the CEFR but by complementing it where necessary, an instrument based on a theoretical framework that would enable test developers and item writers to produce test items that corresponded to the constructs elucidated in the CEFR and that could be calibrated to the CEFR levels. The limited empirical research that we have been able to conduct suggests that, as with other frameworks such as the ACTFL guidelines, the CEFR does not provide sufficient guidance to enable item writers to develop tests at specific levels of the CEFR. However, this tentative conclusion clearly needs to be replicated in much larger studies, which probably can be undertaken only once a body exists of tests and tasks that have been developed explicitly to correspond to the CEFR and that have been empirically linked to the CEFR. Currently, relatively few such tests exist. The CEFR itself is clearly intended more as a user-oriented set of scales than as a constructor-oriented set of scales (Alderson, 1991), and we recommend that in future references to the CEFR, this important distinction be emphasised. The CEFR should not be taken to present a set of specifications for test development at the different levels it posits, but rather it can act, and has indeed so acted within the project reported in this article, as a fruitful starting point for the analysis and development of items and tasks intended to measure reading and listening abilities.

Indications from our necessarily limited research are that the dimensions of the Grid (and thus of the CEFR and its scales) do not alone or maybe even in combination distinguish among the CEFR levels. Indeed, we have proposed an empirical process whereby content analysis of test texts and items should proceed hand in hand with empirical investigations of difficulty and empirical standard-setting procedures.

Nevertheless, lest this conclusion seem unduly pessimistic, we wish to affirm that the project has developed an instrument whose latest version, the Final Grid, provides a promising framework for the characterisation of test items and tasks and thus represents a contribution to the growing literature on the development and use of the CEFR. Hopefully, analysis of the results of further use of the Final Grid will also contribute to a better understanding of what changes as language learners develop in their ability to understand written and spoken texts in a foreign language.

REFERENCES

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

Alderson, J. C. (2000). *Assessing reading*. Cambridge Language Assessment Series. Cambridge, England: Cambridge University Press.

Alderson, J. C. (Ed.). (2002). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies*. Strasbourg, France: Council of Europe.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum Books.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., Tardieu, C. (2004). *Final report of the Dutch CEFR project*. Unpublished manuscript..

Alderson, J. C., and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22,* 301–320.

Allen, E. D., Bernhardt, E. B., Berry, M. T., Demel, M. (1988). Comprehension and text genre: An analysis of secondary school foreign language readers. *Modern Language Journal, 72* (163–172), cited in J. C. Alderson (2000).

Buck, G. (2001). *Assessing listening*. Cambridge Language Assessment Series. Cambridge, England: Cambridge University Press.

Buck, G., Tatsuoka, K., and Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47,* 423–466.

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

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR).*

———*Preliminary pilot version of the manual*. Strasbourg, France: Language Policy Division, Council of Europe.

Council of Europe. (2005). *Reading and listening items*. Strasbourg, France: Council of Europe. Compact disc.

Figueras, N., North, B., Takala, S., Van Avermaet, P., & Verhelst, N. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22,* 261–279.

Gille, E., & Sluiter, S. (2005, June). *The European Anchor Item Bank*. Paper presented at the 2nd annual conference of the European Association for Language Testing and Assessment, Voss, Norway.

Lee, J. F., and Musumeci, D. (1988). On hierarchies of reading skills and text types. *Modern Language Journal, 72,* (173–187), cited in J. C. Alderson (2000).

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

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

Shiotsu, T., and Weir, C. J. (2004). *The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance*. Unpublished manuscript.

APPENDIX A
Selected Sample *Can-Do* Statements from the CEFR for Reading and Listening

A1 Reading
(page numbers refer to CEFR, 2001, English version)

Can understand familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type (p. 24).

I can understand familiar names, words, and very simple sentences, for example on notices and posters or in catalogues (p. 26).

Can understand very short, simple texts a single phrase at a time, picking up familiar names, words, and basic phrases and rereading as required (p. 69).

Can understand short, simple messages on postcards (p. 69).

Can recognise familiar names, words, and very basic phrases on simple notices in the most common everyday situations (p. 70).

Can get an idea of the content of simpler informational material and short simple descriptions, especially if there is visual support (p. 70).

Can follow short, simple written directions, for example, to go from X to Y (p. 71).

C1 Listening

Can understand a wide range of demanding, longer texts, and recognise implicit meaning (p. 24).

Can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. Can understand television programmes and films without too much effort (p. 27).

Can understand enough to follow extended speech on abstract and complex topics beyond his or her own field, although he or she may need to confirm occasional details, especially if the accent is unfamiliar. Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts. Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly (p. 66).

Can easily follow complex interactions between third parties in group discussion and debate, even on abstract, complex unfamiliar topics (p. 66).

Can follow most lectures, discussions, and debates with relative ease (p. 67).

Can extract specific information from poor quality, audibly distorted public announcements, for example, in a station, sports stadium, and so forth (p. 67).

Can understand complex technical information, such as operating instructions or specifications for familiar products and services (p. 67).

Can understand a wide range of recorded and broadcast audio material, including some nonstandard usage, and identify finer points of detail including implicit attitudes and relationships between speakers (p. 68).

Can follow films employing a considerable degree of slang and idiomatic usage (p. 71).

Is skilled at using contextual, grammatical, and lexical cues to infer attitude, mood, and intentions and anticipate what will come next (p. 72).

Can take detailed notes during a lecture on topics in his or her field of interest, recording the information so accurately and so close to the original that the notes could also be useful to other people (p. 96).

Can select an appropriate formulation from a broad range of language to express himself or herself clearly without having to restrict what he or she wants to say (p. 110).

APPENDIX B
Extract from Frame 2 for Reading Level B2

Operation	*What* (*= focus and topic theme*)	*Text*	*Text Features*	*Strategy*	*Conditions and Limitations*
Understand (1)	Relevant details	Correspondence	Related to my field of interest	Reread difficult sections	Difficulty with less common phrases and idioms and with terminology
Scan (2)	Essential meaning	Longer texts, including specialized articles outside my field (1)	Range and type of text only a minor limitation	Can read different types of text at different speeds and in different ways according to purpose and type	With a large degree of independence
Monitor (3)	Information (4)	Highly specialised articles within my field			Dictionary required for more specialized or unfamiliar texts
Obtain (4)	Reference sources	News items (8), articles, and reports on contemporary problems with particular viewpoints			
Select (5)	Content and relevance (8)	Lengthy complex instructions, including conditions and warnings			
Evaluate (6)	Contextual clues	Long and complex texts (2)			
Locate (7)	Ideas and opinions	Contemporary literary prose (1)			
Identify (8)	Both concrete and abstract topics (1)				

APPENDIX C
Final Grid

A. Characteristics of Input Text

1. Text Source: Reading

Abstracts
Advertising material
Blackboard text
Broadcast and recorded spoken text
Brochures
Business letter
Computer screen text
Contracts
Dictionaries
Exercise materials
Guarantees
Instruction manuals
Instructional material
Job description
Journal articles
Junk mail
Labels and packaging
Leaflets, graffiti
Life safety notices
Magazines
Menus
Newspapers
Notices, regulations
Novels
OP text
Personal letters
Programmes
Public announcements, notices
Recipes
Reference books
Regulations
Reports, memorandum
Sacred texts, sermons, hymns
Sign posting
Teletext
Textbooks, readers
Tickets, timetables
Videotext
Visiting cards

2. Text Source: Listening

1. Text source	Debates and discussions (both live and on the media) Entertainment (drama, shows, readings, songs) Interpersonal dialogues and conversations Interviews (both live and on the media) News broadcasts Public announcements and instructions Public speeches, lectures, presentations, sermons Publicity texts (e.g. radio, TV, and supermarkets) Radio phone-in Recorded Tourist information Rituals (ceremonies, formal religious services) Routine commands (instruction/direction by police, customs officials, airline personnel, etc.) Sports commentaries (football, cricket, boxing, horse racing, etc.) Songs and poems Telephone conversations Telephone information (automatic answering devices, weather, traffic conditions, etc.) Traffic information TV, radio documentaries Weather forecasts

Note. (Taken from CEF Table 5 pages 48/9).

2. Authenticity: Reading and Listening

Input text appears to be:

- Genuine
- Adapted/simplified
- Pedagogic

3. Discourse Type: Reading and Listening

Discourse Types		*Examples*
• Mainly argumentative	Comments	By any individual in any situation, pros and cons of an issue, opinions
	Formal argumentation	e.g. Formal debate
• Mainly descriptive	Impressionistic descriptions	e.g. Sports commentaries, physical appearance, layout of room, house, landscape, places
	Technical descriptions	e.g. Presentation of a product
• Mainly expository	Definitions	Brief definitions
	Explications	Broader accounts of (especially) abstract phenomena e.g. lectures, talks
	Outlines	e.g. Programme listings on the radio, time-tables
	Summaries	e.g. An oral account of the plot of a book, summarizing minutes of a meeting
	Interpretations	e.g. Describing a book, an article, etc.
• Mainly instructive	Personal instructions	e.g. Announcements, ads, propaganda, routine commands
• Mainly narrative	Stories, jokes, anecdotes	
	Reports	e.g. News reports, features, documentaries
• Mainly phatic		e.g. Establishing communication, chatting, small talk, etc.

4. Domain: Reading and Listening

Personal: Domain in which the person concerned lives as a private individual, centers on home life with family and friends and engages in individual practices such as reading for pleasure, keeping a personal diary, pursuing a special interest or hobby, etc.

Public: Domain in which the person concerned acts as a member of the general public or of some organization and is engaged in transactions of various kinds for a variety of purposes.

Occupational: Domain in which the person concerned is engaged in his or her profession.

Educational: Domain in which the person concerned is engaged is organized learning, especially but not necessarily within an educational institution.

Note. In many situations, more than one domain may be involved.

5. Topic: Reading and Listening

Select from:

1. Personal identification
2. House and home, environment
3. Daily life
4. Free time, entertainment
5. Travel
6. Relations with other people
7. Health and body care
8. Education
9. Shopping
10. Food and drink
11. Services
12. Places
13. Language
14. Weather
15. Other (please specify)

6. Nature of Content: Reading and Listening

1. Only concrete content
2. Mostly concrete content
3. Fairly abstract content
4. Mainly abstract content

7. Text Length

In words (Reading)
In seconds (Listening)

8. Vocabulary (Reading and Listening)

Select From:

1. Only frequent vocabulary
2. Mostly frequent vocabulary
3. Rather extended
4. Extended

9. Grammar: Reading and Listening

Select From:

1. Only simple structures
2. Mostly simple structures
3. Limited range of complex structures
4. Wide range of complex structures

10. Text Speed: Listening

Select From:

1. Artificially slow
2. Slow
3. Normal
4. Fast

11. Number of Participants: Listening

Select From:

1. One
2. Two
3. More than two

12. Accent/standard: Listening

Select From:

1. Standard accent
2. Slight regional accent
3. Strong regional accent
4. Non-native accent

13. Clarity of Articulation: Listening

Select From:

1. Artificially articulated
2. Clearly articulated
3. Normally articulated
4. Sometimes unclearly articulated

14. How Often Played: Listening

Select From:

1. Played once
2. Played twice
3. Played three times
4. Played more than three

15. Listening / 10. Reading

Comprehensible by Learner at CEFR Level

Below A1
A1
A1/A2
A2
A2/B1
B1
B1/B2
B2
B2/C1
C1
C1/C2
C2
Beyond C2

B. Characteristics of Item

16. Listening / 11. Reading

Item Types

Response Type	Test Method
Selected response	Multiple choice Banked multiple choice True/false Multiple matching Sequencing/ordering jumbled text Citing
Short constructed response	Short answer Cloze (every nth) Gap-filling (one word) C-Test Summary completion Information transfer Sentential response Justify by citing
Extended constructed response (creative, etc.)	Essay Summary Report in own words Justify in own words Other

17. Listening/12. Reading

Operations

Recognise	Main idea/gist	
	Detail	
Make inferences	Opinion	From explicit information
	Speaker’s/writer’s attitude/mood	
	Conclusion	From implicit information
Evaluate	Communicative purpose	
	Text structure/connections between parts	

Item Level Estimated

Please Select;:

Below A1
A1
A1/A2
A2
A2/B1
B1
B1/B2
B2
B2/C1
C1
C1/C2
C2
Beyond C2