Revista de LINGÜÍSTICA Y LENGUAS APLICADAS

Revista de Lingüística y Lenguas Aplicadas
Vol. 10 año 2015, 66-77
EISSN 1886-6298
http://dx.doi.org/10.4995/rlyla.2015.3434

JUDGMENTAL EVALUATION OF THE CEFR BY STAKEHOLDERS IN LANGUAGE TESTING

Mathea Simons
Jozef Colpaert
University of Antwerp (Belgium)

Abstract: *This study provides insights into the judgmental evaluation of the Common European Framework of Reference for Languages (CEFR) by stakeholders or users. Given its widespread use and the debates surrounding it, a deeper analysis was required regarding their experiences when applying the CEFR in their daily practice of language testing, their perceptions on possible improvements and priorities. One hundred eighty-eight users, representing several groups of stakeholders, attended a conference on the topic and participated in discussion groups. These discussion groups were nourished by data obtained by a pre-conference survey and followed by a voting process on priorities for improving the Framework. The results show that the respondents have a positive attitude towards the CEFR. They use it for several purposes and consider its usefulness, authenticity and applicability as positive aspects. The degree of detail and practicality are assessed less positively. The most important recommendation for improvement lies in further fine-tuning and in improving practice and implementation.*

Keywords: *CEFR, language testing, usefulness, improvements, prioritization.*

1. INTRODUCTION

The Common European Framework of Reference for Languages (CEFR), introduced in 2001, has a broad aim and includes language learning, language teaching and language testing. It describes what language learners "have to learn in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively" (Council of Europe, 2001:1). To this end, the CEFR identifies Performance Level Labels, six levels of proficiency (ranging from A1 (lowest) to C2) and Performance Level Descriptions, statements describing what learners can do with language (e.g. CEFR descriptor "Can write simple isolated phrases and sentences" on Performance Level A1) (Papageorgiou, 2010). The Framework aims to enhance transparency and mutual recognition of qualifications by providing an explicit description of objectives, content and methods as well as a by giving objective criteria for describing language proficiency (Council of Europe, 2001).

In the field of language testing, the CEFR has gradually been adopted and is known nowadays as an important instrument. Indeed, Little (2007) states that the impact of the CEFR on language testing by far outweighs its impact on curriculum design and pedagogy.

On the one hand, the CEFR has been used as a reference point for the development of tests and assessment procedures, for example, with DIALANG at the European level as well as with other tests in national and regional educational settings (e.g. Alderson & Huhta, 2005; Little, 2005). It also has been used as a common set of standards for (existing) examinations. The Framework has increasingly been referred to as "the basis for the mutual recognition of language qualifications in Europe" (Figueras *et al.*, 2005:65).

On the other hand, since its introduction, the CEFR has not been without criticism. It has been the subject of discussions along the way, and several issues have been raised. Some authors question the overall usability of the CEFR for language testing, stating that it is "not sufficiently comprehensive, coherent or transparent for uncritical use in language testing" (Weir, 2005:281). Other colleagues claim that the CEFR can only serve as a starting point (Alderson *et al.*, 2004 in Little, 2007). The impact of the political agenda on language testing has also been subject to criticism because institutions in various countries are being encouraged and funded to achieve harmonization with a view to European Citizenship (Fulcher, 2004).

Received: 2014/12/03 Accepted: 205/02/17

In order to identify the main points of criticism, we conducted a review of the literature regarding the CEFR. Generally speaking, the elements addressed in the retrieved manuscripts can be related to the content and structure of the Framework, its usefulness for language testing and/or its actual usage or misuse (Keddle, 2004; North, 2007). More specifically, five categories can be distinguished. A first point of criticism is the low *degree of detail* of the Framework: the CEFR does not always include what stakeholders need. There are important gaps in the CEFR scales (Weir, 2005; Alderson, 2007; Fulcher, 2010) and language testers are facing the difficulty of having to define their own concepts as well as to develop their own reporting scales (Alderson & Huhta, 2005).

A second element is the *clarity* of the Framework. The CEFR is not always easy to understand. As the CEFR is the result of a compromise between several traditions, it appears to be a complex document. It encompasses a vast array of conceptual domains (Picardo, 2011). There is a lack of definitions and, in addition, there are ambiguities and inconsistencies in the use of terminology (Alderson, 2007). Fulcher (2004) states that the meta-objective of providing proficiency descriptors that are applicable across languages, requires a framework so abstract that it is not a framework any more, but a model. Thus, in order to maintain its applicability across languages, the CEFR in fact becomes less precise and therefore more difficult to follow.

A third element is a problem of *authenticity*, that is, the fact that there is insufficient correlation with real-world language use tasks. The cause of this problem can be found within the scope of what the CEFR attempts, largely due to the fact that “the illustrative scales are only partly validated by empirical research” (Little, 2007:648). The possible impact of variation in terms of contextual parameters on the actual difficulty level of can-do statements has not been sufficiently taken into account (Weir, 2005), which reduces the relevance of the CEFR to the varied contexts of real-world situations.

A fourth element relates to the *applicability* of the CEFR. Is it possible to use the Framework as such in the field of language testing? Several problems have been reported on this topic. Identifying separate CEFR performance level labels seems to be 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 (Figueras *et al.*, 2005; Alderson *et al.*, 2006; Figueras, 2012). The generalizability of the level labels has been pointed out as a problem. Hulstijn (2007), for example, criticizes the overall scales as defined in the Framework, arguing that the CEFR scales do not take into account all types of second language users, which complicates its use in the field of language testing. Furthermore, he states that the aspect of applicability is intertwined with issues of quantity and quality.

A fifth element is its *practicality*. How easy is it to build CEFR-based tests? This element addresses both the problems of test development and the actual linking of existing tests to the CEFR. Authors report the difficulty of confirming whether or not tests are actually measuring the CEFR levels they claim they are measuring (Papageorgiou, 2010). As Figueras *et al.* (2005:264-265) point out, “as useful as a qualitative grid may be in developing curricula and in formative assessment (…), for summative assessment, as in examinations, some form of quantification is needed”.

The general *usefulness* of the Framework continues to be questioned. Does the CEFR help stakeholders in the field of language testing in their job? Figueras *et al.* (2005:270) state that it is still an abstract descriptive system leaving a “considerable gap between the description available and the practical needs of both large-scale and small-scale assessment”. Other authors mention problems for specific groups of users, for example, for school-based teachers (Keddle, 2004; Picardo, 2013). Its organisational complexity also can prove to be intimidating for some practitioners (Picardo, 2011), and other authors mention issues on the use of the CEFR as a policy document and as a marketing and recognition tool for assessments (Fulcher 2004a, 2004b, 2009; Shohamy and McNamara, 2009).

We can conclude that the CEFR is a useful reference tool in the field of language education and language testing. Nevertheless, it has been criticised in the field in which it has been particularly influential, namely language testing and assessment.

2. RESEARCH QUESTIONS

Taking into account the importance of the CEFR in the field of language testing on the one hand and the identified issues on the other, it is important to know how actual stakeholders of the CEFR feel about the Framework, how they view its usefulness, and where they see a need for clarification and improvement. In the light of this context, the authors of this article, both involved in test development, teacher training and in the organization of national and international conferences, decided to gauge the perceptions of stakeholders i.e. persons involved in the field of language testing such as language teachers, textbook authors, curriculum developers, policy makers, researchers, language testing associations, publishers, students and instructional designers, hereafter referred to as ‘users’. The following research questions were formulated:

RQ1 How useful is the CEFR in concrete language testing situations, taking into account the issues mentioned in the literature?

RQ2 Which aspects of the CEFR are amenable to improvement according to users of the CEFR in the field of language testing?

RQ3 Which points for improvement should be handled first?

The main purpose of this study was to shed light on the perceptions of users in order to inform policy makers and researchers how the CEFR is *perceived* and how it can be improved. The results are intended as indications, which, combined with others, should give direction to future policy and research on the CEFR.

3. RESEARCH DESIGN

3.1 Data collection

In order to formulate answers to these questions, we decided to follow a staged approach as we considered the answers to RQ1 as a possible basis for answering RQ2 and RQ3. We first designed a short online survey, taking into account RQ1. In order to answer RQ2 and RQ3, we decided to adopt a qualitative approach (i.e. discussion groups). We brought together a group of users of the CEFR in the field of language testing and assessment during an international conference on the topic. This conference, entitled ‘Language testing in Europe. Time for a new framework?’ took place at the University of Antwerp (Belgium), 27-29 May 2013.

The pre-conference survey consisted of five questions (see Appendix 1), preceded by an introduction in which the context of the questionnaire, (i.e. language testing), was clarified. In order to maximize the number of responses, we opted for a short survey, including only three questions on the familiarity with and the effective use of the CEFR, and two questions evaluating the issues found in the literature. Each question also offered an ‘Other’ category, which allowed participants to submit open-ended answers and comments. The data obtained by the online survey (RQ1) were presented at the beginning of the conference. During the conference, the respondents were invited to work together in discussion groups (groups of 2 to 4 people). The groups were asked to write at least two or three tweet-like recommendations for improvement (+/- 140 characters) (RQ2). These recommendations were presented to the respondents during the closing session of the conference, and voted upon according to their perceived priority by means of the voting system *PowerVote* (RQ3), a voting application that makes it possible to gather data and to display them instantly. The respondents received an individual voting device in order to communicate their personal prioritization.

3.2 Respondents

One hundred and eighty-eight users of the CEFR (Nman=61; Nwoman=127) participated in the study. They attended the international conference on the topic, participated in the discussion groups as well as in the final voting on priorities. Twenty-six countries were represented, from which eight were non-EU countries. The respondents were invited to give information on their current position. Table 1 gives an overview of the positions mentioned by the respondents. It is important to clarify that several participants indicated a combination of two or more functions (e.g. language teacher in higher education, involved in research and in test development). Therefore the total number of respondents exceeds the number of participants in the conference (N=188).

Table 1. Current position of the respondents.

Function	N
- Language teacher (secondary education)	18
- Language teacher (adult education)	36
- Language teacher (higher education)	54
- Researcher (PhD, Researcher)	61
- Policy maker (Directors, Coordinators, national or international organizations)	49
- Test developer	17
- Counselor/coach (Staff, Students)	12
- Teacher trainer	3
- Publisher	2

The online survey, administered in order to answer RQ1, was filled in by 115 respondents (64,7%) before the actual start of the conference. The data give an idea of the familiarity with and the effective use of the CEFR. Table 2 shows that the respondents are familiar with the Framework.

Table 2. Current use of the Framework (a).

In my job I *mostly* use the CEFR as follows: *(one possible answer)*	**N=115**	**%**
- I never use the CEFR	8	7%
- I work with my global idea of the CEFR	13	11.4%
- I use the general CEFR labels	27	23.7%
- I use the detailed descriptors	42	36.8%
- I use a national or regional specification of the CEFR	14	12.3%
- I use my own/ my institution's specification of the CEFR	10	8.8%

Of the respondents, 7% stated they never use the CEFR. This result can be explained by the fact that some user groups (e.g. policy makers, publishers) never actually use the Framework as a reference tool, although they know the concept and content. The data also show that the respondents use the Framework in varying degrees of detail (ranging from concrete descriptors to general concepts) and adaptation (ranging from literal application to highly adapted to the local situation). Almost 37% of the respondents said they used the Performance Level Descriptions, while 23.7% used the more general Performance Level Labels. One respondent in five stated that they were using an adapted version of the Framework.

The respondents used the CEFR for all mentioned purposes. Although the frequencies of these purposes are well distributed, Table 3 shows that the most important application of the CEFR was for designing language tests that corresponded to CEFR levels (58.7%). This purpose was followed by two other activities, that is, informing the content of a teaching syllabus or curriculum (49.5%) and designing teaching or learning tasks and activities (46.8%). The alignment of existing materials to the CEFR was clearly less important (30.3%). The open-ended 'Other' option was mainly selected by participants to focus on test development for diagnostic purposes, self-evaluation, or raising awareness.

Table 3. Current use of the Framework (b).

When I use the CEFR in my job, I use it to: *(more than on possible answers)*	**N=115**	**%**
- inform the content of a teaching syllabus/curriculum	57	49.5%
- inform/train teachers about CEFR levels	53	45.9%
- design teaching/learning tasks and activities	54	46.8%
- align existing teaching tasks to the CEFR	32	27.5%
- design tests that correspond to the CEFR	67	58.7%
- align existing tests to the CEFR	35	30.3%

As far as the reason why they use the CEFR is concerned (see Table 4), the respondents indicated that they used it far less because colleagues (4.8%) or superiors (10.5%) asked them to do so, but mainly because research studies indicate that the CEFR is important (59%) or because their institution required them to do so (56.2%).

Table 4. Current use of the Framework (c).

When I use the CEFR in my job, I do this because: *(more than on possible answers)*	**N=115**	**%**
- the institution I work for requires me to do this.	65	56.2%
- my immediate supervisor expects me to do this.	12	10.5%
- colleagues tell me the CEFR is important.	6	4.8%
- research studies I have read convince me the CEFR is important.	68	59%

The open-ended 'Other' option was mainly selected by participants to repeat their interest in the CEFR itself as a reason for using it, because of its inherent qualities such as usability and applicability. Its usefulness is being made more explicit as a communication tool among teachers, as a certification, benchmarking and standardization tool, or as a tool for designing tests. The expectations of external partners were also often mentioned as an important reason for using the CEFR. Finally, some participants stated that they use the CEFR simply because there is no alternative.

3.3 Data analysis

The data obtained by means of the online pre-conference survey were analysed using descriptive statistics (RQ1). The data of the discussion groups consisted of recommendations, formulated as tweet-like suggestions for improvement (+/- 150 characters). These data were transcribed verbatim. The first phase consisted of an explorative, detailed analysis in order to define key concepts in the data. In the second phase, the data were entered in NVivo10 software, analysed again and coded manually. During this coding process, individual recommendations were selected and coded using *nodes*. A node is a collection of references about a specific theme. Examples of nodes used in the coding process were 'gaps', 'definitions', 'forum' etc. Thanks to this procedure, one single recommendation can be coded using one or more nodes which prevents data loss. All the recommendations about a particular theme or topic were gathered into a node for further exploration allowing at the same time to quantify the recommendations (RQ2). The results of the final voting on priorities by means of the PowerVote Software were analysed descriptively (RQ3).

4. RESULTS

4.1 Usefulness of the CEFR for language testing (RQ1)

The aim of the pre-conference survey was to shed light on the importance of the (problematic) issues described in the literature which possibly complicate the use of the CEFR in the field of language testing. Table 5 shows that usefulness ("Does it help me in my job?"), authenticity ("Does it correlate with real-world language use tasks?") and applicability ("Can I use it as such in my situation?") of the CEFR as a whole were being perceived as positive.

Table 5. Global evaluation of the CEFR.

I evaluate the CEFR on the following points as:					
(N=115)	**- -**	**-**	**+**	**+ +**	**No opinion**
- Degree of detail (= Does the Framework include what I need?)	4.6%	36.7%	45.9%	8.3%	4.5%
- Clarity (= Is is easy to understand/ remember?)	0.9%	33%	53.2%	11%	1.9%
- Authenticity (= Does it correlate with real-world tasks?)	0%	13.8%	61.5%	22%	2.8%
- Applicability (= Can I use it as such in my situation?)	0%	16.2%	57.7%	23.4%	2.7%
- Practicality (= How easy is it to make CEFR based tests?)	0.9%	24.5%	56.7%	0.9%	17%
- Usefulness (= Does it help me in my job?)	0%	6.5%	50.5%	35.3%	7.7%

More than 85% of the respondents state that the CEFR helps them in their job and they evaluate the global usefulness as rather positive or very positive. Another aspect which is positively rated is authenticity, with 61.5% of the respondents evaluating it as positive, 22% as very positive. A similar result is observed for applicability, with more than 80% of the respondents evaluating it as (very) positive.

The result for clarity was less explicit: 64% of the respondents evaluate this aspect as positive, 34% as negative. Practicality and Degree of detail were perceived as a little less positive. The data also show that respondents hesitated in their evaluation of these aspects, as 4.5% (for Degree of detail) and 17% (for Practicality) did not have a clear opinion. Degree of detail was evaluated less positively. It was the only aspect with a substantial group of respondents evaluating it as 'very negative' (4.5%)

The open-ended 'Other' option was mainly selected to insist on the fact that the CEFR is a useful framework but that it should be specified in more detail in terms of linguistic aspects. Differences between adults and young learners and differences between countries were emphasized, together with the fact that the implementation of the CEFR entails a huge investment in resources and time.

Next to an evaluation of the CEFR as a whole, the respondents were invited to evaluate the Performance Level Labels and the Performance Level Descriptions regarding their definition and the difficulty of the tasks. Table 6 gives an overview of the answers:

Table 6. Evaluation of the Level Labels and Level Descriptions.

I evaluate the CEFR labels and descriptions as: (N=115)	**- -**	**-**	**+**	**+ +**	**No opinion**
- Difficulty of the tasks described in the CEFR labels	0%	12%	60.2%	11.1%	16.7%
- Definition of the CEFR labels	0%	16,7%	60.2%	13.0%	10.2%
- Difficulty of the tasks described in the descriptions (i.e. can do statements)	0%	12,1%	60.7%	14.0%	13.1%
- Definition of the descriptions (i.e. can-do statements)	0%	15%	59.8%	15.9%	9.3%

The data reveal that the definition of the Level Labels and the Level Descriptions, as well as the difficulty of the tasks described in it, were evaluated positively. No extreme negative position ('very negative') could be found. At the same time, the data show that respondents hesitated about the evaluation they gave to the CEFR labels and descriptions as more than 10% of the respondents had no opinion on each of the aspects.

The open-ended 'Other' option reflected largely individual differences in the perception of the CEFR level labels and descriptors, but recurring themes were the lack of lexico-grammatical specifications, the academic and ill-informed nature of the C levels and the lack of consistency in the interpretation of the descriptors.

In conclusion, the perceptions of and attitudes towards the CEFR were rather positive. The attitudes of participants can even be called constructive and committed: they tried to identify concrete aspects which were amenable to improvement. They did note, however, that it still requires a considerable amount of time and energy in order to understand and to use the labels and the descriptions, due to their lack of practicality and degree of detail, especially regarding their lexico-grammatical specification. Participants seem to feel that it is a common responsibility to continue to build this CEFR as a tool for internationalization, communication, standardization and design.

4.2 Aspects amenable to improvement (RQ2)

As a result of the discussion groups (each consisting of 2 to 4 participants) organized during the conference, we received 151 recommendations. These recommendations were formulated as tweet-like suggestions for improvement (max. 140 characters). Some examples of the raw data include: "Fill in gaps for missing descriptors without becoming too prescriptive", "Add more descriptors on pronunciation (showcase)", "Make the wording of descriptors more consistent and remove inconsistencies", and "Descriptors should be provided for/adapted to LSP (language for specific purposes) settings".

These recommendations were read carefully and categorized using NVivo nodes (e.g. specific groups of learners, descriptors, gaps), thus allowing the researchers to reduce and quantify the number of recommendations. In this way, 32 recommendations could be distinguished. A summary is included in Appendix 2.

The data show that there is a great deal of diversity in the recommendations. Nevertheless, it is possible to categorize them into three main topics (see Appendix 2):

1. *Fine-tuning*: On the whole, 15 global recommendations, representing 63 specific recommendations (=41.7%) mention fine-tuning as a genuine concern. This concern also confirms the results of the pre-conference survey (cf. supra), which state that the degree of detail of the Framework does not entirely meet the expectations of the users. Fine-tuning is required as far as missing descriptors or incomplete templates are concerned. While doing this, the latest findings in research can be taken into account and vagueness of terminology can be avoided.

2. *Improving practice and implementation*: Eleven global recommendations, representing 54 specific recommendations (=35.7%) underline the need to promote and facilitate good practice. The field asks for better structured information, more exchange and information. They also emphasize the need for some control over the use that is made of the Framework in real educational settings. By doing so, stakeholders should be involved. This way, expectations could be managed in a better way and critical awareness would be raised in a more significant manner.

3. *Extensions:* Six global recommendations, representing 34 specific recommendations (=22.5%) ask extensions to other groups of learners, other contexts, new skills and knowledge or specific fields. As far as groups of learners are concerned, stakeholders ask more attention for specific professional contexts as to the age and the mother tongue(s) of learners. The lowest and highest levels of the CEFR (and beyond) also should be worked out in more detail. As far as skills and knowledge are concerned, the field requires

more information to assess specific skills (e.g. listening) and specific knowledge (e.g. lexis, grammar). Special attention is required for phonological control and pronunciation.

The following table provides an overview of the most frequently suggested recommendations, indicating how often they were mentioned by the participants.

Table 7. Most frequently suggested recommendations for improvement.

Recommendation	N
- Provide objective and well defined criteria/ clearer/ more consistent descriptors	15
- Adapt to specific groups of learners (e.g. young learners)	12
- Provide a platform for the exchange of good practices, examples and evidence	12
- Adapt to specific contexts: different professional contexts and specific purposes (e.g. Academic English)	11
- Provide more examples for course designers/ teachers	11
- Explain how transitions are to be made from one level to another	7
- Raise critical awareness of actors ; manage expectations	7
- Fill the gaps in the CEFR/ missing descriptors	6

These recommendations were followed by seven other recommendations, each of them provided by five groups of respondents. These included making the CEFR more user-friendly, providing more descriptors on phonological control and pronunciation, and providing better structured information on the official website. We chose to set this group aside and to investigate more deeply the eight most frequently suggested recommendations.

4.3 Prioritization of recommendations for improvement (RQ3)

During the closing session of the conference, the list presented in Table 7 was submitted to the respondents. This complementary step was taken in order to define priorities. They were asked to determine the three main actions to be undertaken. The PowerVote system was used again to such effect. Respondents were asked to state which of the actions were to be given priority: 1 (most important), 2 or 3. Table 8 gives an overview of the percentages given to each recommendation.

In order to establish a top 3 list of recommendations, taking into account the individual priorities of each recommendation, we calculated scores as follows: 3 for priority 1; 2 for priority 2 and 1 for priority 3. The following results were thus obtained:

Table 8. Priorities for improvement.

	Priority 1	Priority 2	Priority 3	**Calibrated total**	**Global priority**
Provide objective and well defined criteria/ clearer/ more consistent descriptors	17%	17%	20%	**105**	3
Adapt to specific groups of learners (e.g. young learners)	5%	2%	4%	**23**	
Adapt to specific contexts: different professional contexts and specific purposes (e.g. academic English)	6%	5%	7%	**35**	
Provide more examples for course designers/teachers	9%	12%	18%	**69**	
Provide a platform for the exchange of good practices, examples and evidence	19%	30%	9%	**126**	1
Raise critical awareness of all stakeholders; manage expectations	22%	13%	16%	**108**	2
Explain how transitions are to be made from one level to another	5%	11%	15%	**52**	
Fill the gaps in the CEFR/ missing descriptors	17%	10%	11%	**82**	

Table 8 confirms the hypothesis that users primarily asked for improvements in the field of practice and implementation as well as a fine-tuning of the current framework. The respondents considered as priority 1 "*Provide a platform for the exchange of best practices, examples and evidence".* They demanded more exchange and wished that more information could be given by means of concrete examples and best practices both for test developers and for teachers. They also believed that evidence from research should be made public. As to who the main actors of the process should be, the results pointed towards Europe and the universities.

"*Raise critical awareness of all stakeholders; manage expectations"* was identified as being the second priority. The CEFR is merely an instrument which should be adapted to the specific educational setting it will be used in.

The actual use of the CEFR shows that stakeholders in the field sometimes interpreted this use in a different way, that is, without fine-tuning the (general) framework to the specific needs and characteristics of the settings they work in. This way, exaggerated and even erroneous expectations are being created. These expectations may distort or bias the stakeholders' judgments. Therefore stakeholders might interpret and evaluate the Framework in a way which does not correspond to its initial and primary goal. In order to avoid these problems, the respondents underlined the importance of critical awareness among stakeholders. Expectations should be managed and there is a need to monitor the actual use of the Framework in real educational settings. Priority 3 is "*Provide objective and well-defined criteria/ clearer/ more consistent descriptors".* Respondents confirmed that some descriptors are too vague, thus complicating the development and the actual linking of language tests to the CEFR. Although this priority is very clear, it is less obvious how researchers and other stakeholders in the field can contribute to this priority.

5. CONCLUSION AND DISCUSSION

The introduction of the CEFR was a milestone in the field of language education, especially in the field of language testing. It is being used as a reference tool for the development of language tests as well as for the linking to existing tests. However, it has also received a fair amount of criticism, and, interestingly, its value and usefulness has probably been most questioned in the field in which it has been particularly influential, namely language testing and assessment. Taking into account the widespread use of the Framework and the debates surrounding it, it is important to know how users of the CEFR feel about the Framework, how they view its usefulness, and where they see a need for improvement, clarification and argumentation.

This article reports on a study in which a large group of users of the CEFR, representing several groups of stakeholders, were asked to assess the usefulness and the usage of the Framework. They were also invited to make recommendations for improvement and to identify priorities. The study was mainly qualitative in nature. One hundred eighty-eight conference participants, all users of the CEFR, were asked to discuss their perceptions of the CEFR, the points in the CEFR that need improvement and a subsequent prioritization of the most often mentioned points for improvement. The respondents worked in discussion groups of two to four persons. A pre-conference survey was used to explore familiarity with, effective use and evaluate issues found in the literature.

The results showed that the respondents had a positive attitude towards the Framework and give evidence of a constructive insight. Usefulness, authenticity and applicability were assessed positively. Degree of detail and practicality were seen less positively. As far as global recommendations for improvement are concerned, the respondents primarily asked for fine-tuning. There is a need for more detailed descriptors and for filling the gaps. A second recommendation concerns the improvement of practice and implementation. This can be accomplished by providing a platform for exchange as well as by managing expectations. A third recommendation relates to the extension of the Framework to specific groups of leaners as well as to specific contexts. In a last research step, respondents were invited to vote upon specific suggestions, which enabled us to identify three priorities: (1) Provide a platform for the exchange of good practices, examples and evidence; (2) Raise critical awareness of all stakeholders and manage expectations, and (3) Provide objective and well defined criteria/ clearer/ more consistent descriptors.

As far as the recommendations for improvement are concerned, several of the topics can be found in the literature (e.g. lack of clarity in the existing descriptors, need for more domain-specific descriptors, need for descriptors that are relevant to specific test taker age groups, need for more descriptors for higher levels). This study made it possible to have a clearer view on the importance of each individual issues and on the priorities for the future. The prioritization of the areas related to the CEFR might surprise: the top two concerns of the respondents are that a platform is needed (for exchanging good practices, examples of evidence to support the use of the Framework) and that the stakeholders' awareness of the CEFR – its potential as well as its limitations and issues – should be increased. While these points are by no means unknown to people working with the Framework, they have not been given the attention they apparently deserve.

The research design applied in this study clearly has its limitations. The research questions were answered by a specific group of respondents, that is, the attendees of a conference on the topic. This implies that there is a certain weight of participants, but that the representativeness of the findings cannot be guaranteed. On the other hand, the respondents represent groups of users (researchers, practitioners and policy makers) who regularly work with the CEFR and therefore have a clear view of the problems users are facing.

The objective of this study was to shed a light on usage and usefulness of the CEFR in the field of language testing, and more particularly on the points of the Framework that need improvement and the prioritization. The respondents of this study were users of the CEFR. We collected information about their current position. Due to the fact that several respondents mentioned more than one occupation it was not possible to see if, for example, respondents specializing in language testing had different views from SLA researchers or teacher educators or language teachers. A more in-depth study, taking into account the statuses of the respondents as well as their motivations and degrees of involvement in language testing could give a more accurate image and a clearer insight in priorities for specific groups of users.

The current study gives an overview of the usefulness of the CEFR in the field of language testing, the aspects amenable to improvement and the priorities. As far as the aspects amenable to improvement are concerned, respondents pointed out global ideas (i.e. improving practice and implementation) as well as specific priorities (i.e. fine-tuning). These specific recommendations could be the focus of a more in-depth study in order to identify the respective descriptors, templates or terminology by means of specific questionnaires or discourse-based research such as interviews or focus groups.

The study revealed that the vertical dimension (i.e. the Level Labels and Descriptors) of the CEFR is the more problematic one. The vertical dimension is the most influential part of the CEFR. Nevertheless, this does not imply that the horizontal dimension (i.e. the content categories) (cf. Alderson, 2005; Kaftandjieva, 2007) should be neglected. Future study differentiating between both dimensions could shed valuable light on the aspects of the CEFR that stakeholders believe could be further developed. Users of the CEFR certainly do appear to see a role for it in language testing, but more work needs to be done to determine its strengths and shortcomings, which can lead to more informed decisions regarding its use.

REFERENCES

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

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*/3, 301-320. http://dx.doi.org/10.1191/0265532205lt310oa

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. and Tardieu, C. (2006). “Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project”, *Language Assessment Quarterly, 3*/1, 3-30. http://dx.doi.org/10.1207/s15434311laq0301_2

Alderson, J.C. (2007). “The CEFR and the need for more research”, *The Modern Language Journal, 91*/4, 659–663. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_4.x

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

Figueras, N., North, B., Takala, S., Verhelst, N. and Van Avermaet, P. (2005). “Relating examinations to the Common European Framework: a manual”, *Language Testing, 22*/3, 261-279. http://dx.doi.org/10.1191/0265532205lt308oa

Figueras, N. (2012). “The impact of the CEFR”, *ELT Journal, 66*/4, 477-485. http://dx.doi.org/10.1093/elt/ccs037

Fulcher, G. (2004a). “Are Europe’s tests being built on an ‘unsafe’ framework?”. Retrieved 11/11/2014 from http://education.guardian.co.uk/tefl/story/0,5500,1170569,00.html

Fulcher, G. (2004b). “Deluded by artifices? The Common European Framework and harmonization”, *Language Assessment Quarterly, 1*/4, 253-266. http://dx.doi.org/10.1207/s15434311laq0104_4

Fulcher, G. (2009). "*Test use and political philosophy". Annual Review of Applied Linguistics, 29,* 3-20. http://dx.doi.org/10.1017/S0267190509090023

Fulcher, G. (2010). “The reification of the Common European Framework of Reference (CEFR) and effect-driven testing”, *Advances in research on language acquisition and Teaching: Selected Papers*, 15-26.

Hulstijn, J.H. (2007). “The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency”, *The Modern Language Journal, 91*/4, 663–667. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_5.x

Kaftandjieva, F. (2007). “Quantifying the quality of linkage between language examinations and the CEFR”, in C. Carlsen and E. Moe (eds.) *A human touch to language testing.* Oslo: Novus Press, 33-43.

Keddle, J.S. (2004). “Insights from the Common European Framework, 2004”, *The CES and secondary school syllabus*. Oxford: Oxford University Press, 43-54.

Little, D. (2005). “The Common European Framework and the European Language Portfolio: involving learners and their judgements in the assessment process”, *Language Testing, 22*/3, 321-336. http://dx.doi.org/10.1191/0265532205lt311oa

Little, D. (2007). “The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy”, *The Modern Language Journal, 91*/4, 645-655. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_2.x

North, B. (2007). “The CEFR illustrative descriptor scales”, *The Modern Language Journal, 91*/4, 656-659. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_3.x

Papageorgiou, S. (2010). “Investigating the decision-making process of standard setting participants”, *Language Testing, 27*/2, 261-282. http://dx.doi.org/10.1177/0265532209349472

Picardo, E. (2011). “Du CECR au développement professionnel : pour une démarche stratégique”, *Revue canadienne de linguistique appliquée,* 14/2, 20-52.

Picardo, E. (2013). “(Re)conceptualiser l’enseignement d’une langue seconde à l’aide d’outils d’évaluations:comment les enseignants canadiens perçoivent le CECR”, *Canadian Modern Language Review*, 386-414. http://dx.doi.org/10.3138/cmlr.1737.386

Shohamy, E. and McNamara, T. (2009). “Language tests for citizenship, immigration, and asylum”, *Language Assessment Quarterly, 6*/1, 1-5. http://dx.doi.org/10.1080/15434300802606440

Urkun, Z. (2008). “Re-evaluating the CEFR: an attempt”, In *The Common European Framework of Reference for Languages (CEFRL): Benefits and Limitations.* IATEFL TEA SIG Croatia Conference Proceedings, pp. 10-14, Canterbury.

Weir, C. (2005). “Limitations of the Common European Framework for developing comparable examinations and tests”, *Language Testing, 22*/3, 281-300. http://dx.doi.org/10.1191/0265532205lt309oa

APPENDIX 1 – PRE CONFERENCE SURVEY

1. In my job I mainly use the CEFR as follows: (*one possible answer*)

O I never use the CEFR
O I work with my global idea of the CEFR
O I use the general CEFR labels
O I use the detailed descriptors
O I use a national or regional specification of the CEFR
O I use my own/ my institution's specification of the CEFR

2. When I use the CEFR in my job, I use it to: (*more than one possible answers*)

O inform the content of a teaching syllabus/curriculum
O inform/train teachers about the CEFR
O design teaching/learning tasks and activities that correspond to the CEFR
O align existing teaching tasks to the CEF
O design tests that correspond to the CEFR
O align existing tests to the CEFR

3. When I use the CEFR in my job, I do this because: (*more than one possible answers*)

O the institution I work for requires me to do this.
O my immediate superior expects me to do this.
O colleagues tell me the CEFR is important.
O research studies I have read convince me the CEFR is important.

4. I evaluate the CEFR on the following points as:

	--	-	+	++	**No opinion**
Degree of detail (= Does it include what I need?)					
Clarity (= It is easy to understand/ remember?)					
Authenticity (= Does it correlate with real-world language use tasks?)					
Applicability (= Can I use it as such in my situation?)					
Practicality (= How easy is it to make CEFR based tests?)					
Usefulness (= Does it help me in my job?)					

5. I evaluate the CEFR labels and descriptions as:

	--	-	+	++	**No opinion**
Difficulty of the tasks described in the CEFR labels					
Definition of the CEFR labels					
Difficulty of the tasks described in the descriptions (i.e.'can do statements')					
Definition of the descriptions (i.e. 'can do statements')					

APPENDIX 2 – RECOMMENDATIONS FOR IMPROVEMENT

Recommendation n = 151	**Number***	**Category****
Provide objective and well defined criteria/ clearer/ more consistent descriptors	15	F
Adapt to specific groups of learners (e.g. young learners)	12	E
Provide a platform for the exchange of good practices, examples and evidence	12	I
Adapt to specific contexts: different professional contexts and specific purposes (e.g. Academic English)	11	E
Provide more examples for course designers/ teachers	11	I
Explain how transitions are to be made from one level to another	7	I
Raise critical awareness of actors ; manage expectations	7	I
Fill the gaps in the CEFR/ missing descriptors	6	F
Define in more detail specific skills (e.g. listening)	5	F
Define the used terminology (e.g. verbs, adjectives, adverbs); avoid vagueness	5	F
Describe knowledge (lexis, grammar) in more detail	5	F
Elaborate on the lowest/highest levels and beyond	5	F
Involve all possible stakeholders when defining the descriptors (learners, teachers, employers etc.)	5	I
Make the CEFR more user-friendly ; provide an all-in-one system	5	F
Provide better structured information on the official website (for all target groups)	5	I
Provide more descriptors on phonological control and pronunciation	5	E
Adapt to 21st century knowledge (new semantic fields; text types)	3	F
Adapt to 21st century skills	3	F
Control the use that is made of the CEFR in real educational contexts	3	I
Provide a revised version based on the latest findings	3	F
Provide language-specific descriptors	3	E
Describe vocabulary and grammar discretely	2	F
Separate linguistic proficiency and socio-/intercultural skills	2	F
Take into account multi- and plurilingualism	2	E
Templates should combine formative and summative aspects	2	F
Add a glossary that could be used at the same time as an analytical index	1	E
Deal better with balance between focus on form and communicative aspects	1	F
Ensure that all products stay public domain (even commercial-owned improvements)	1	I
Give and share practical developments for all languages, not only for English	1	I
Increase assessment and text development literacy across the language teaching industry	1	I
Make globalization work: more collaboration between SLA and Language testing; take out the word 'European'	1	I
Templates should combine analytic and holistic assessment	1	F

* Number of times the recommendation was made.
** F= fine-tuning; E = extensions; I = improving practice and implementation