Behavioural scales of language proficiency: Using the Common European Framework of Reference

Spyros Papageorgiou

University of Michigan

Abstract

The advent of the Common European Framework of Reference has resulted in increased interest in behavioural scales of language proficiency. However, critics have raised concerns as to the limitations of the CEFR for presenting an acquisitional hierarchy based on theories of Second Language Acquisition.

This paper examines how Framework users perceive language development in the CEFR scales. The data originates from a project on relating exams to the CEFR and was analysed using many-facet Rasch measurement with 12 project participants. Aspects of language behaviour in the CEFR scales that participants found hard to interpret are discussed and suggestions for further research from both SLA and language testing perspectives are made in order to understand the characteristics of language development at different CEFR levels. This is of interest to researchers in a number of areas, given the extensive use of the CEFR in the fields of language teaching, learning and assessment.

Keywords: language assessment, CEFR, SLA, language proficiency scales

1. Introduction

The advent of the Common European Framework of Reference (CEFR, Council of Europe 2001) has resulted in increased interest in the development and use of behavioural scales of language proficiency. It is generally accepted that despite the richness of the CEFR volume regarding second language learning, its scaled descriptors are by far the most well-known part (North 2005). The development of the CEFR scales is the result of extensive research (North 2000, North and Schneider 1998); however its use has not been without problems. Relevant studies have shown that researchers found difficulties when using the CEFR to

- design test specifications (Alderson *et al.* 2006)
- measure progression in grammar (Keddle 2004)
- describe the construct of vocabulary(Huhta and Figueras 2004)
- design proficiency scales (Generalitat de Catalunya 2006)

Critics have also raised concerns as to the limitations of the CEFR for comparing language qualifications and for presenting an acquisitional hierarchy based on theories

of Second Language Acquisition (Fulcher 2004a, 2004b, Weir 2005). Calls for collaboration between SLA and language testing researchers point out the benefit of such joint research in better understanding the language development represented by behavioural scales (Brindley 1998) and in particular the CEFR scales (Alderson 2005a, 2005b).

Within this context, the study reported in this paper looks at the process of relating language examinations to the CEFR and in particular the use of the CEFR scales by participants in this linking process. Three research questions are addressed:

1. Can users of the CEFR rank-order the scaled descriptors from lower to higher levels in the way they appear in the 2001 volume?
2. If differences in rank-ordering exist between the users of the CEFR and the 2001 volume, why does this happen?
3. Can training contribute to more successful scaling?

Given the increased reference to the CEFR by examination providers, who often claim that their examinations are linked to the CEFR levels, the above questions are crucial for the validity of such claims. This is because the linking process, presented in the Council of Europe's Manual (Council of Europe 2003, Figueras *et al.* 2005), is based on judgements of participants who have been previously trained in using the CEFR and have a good understanding of its scales. If participants cannot rank-order the descriptors in the intended order (i.e., from lower to higher levels), then their understanding of the scales is probably limited and any claim as to how a language examination is linked to the CEFR should be disputed.

2. Background to the study

The study reported here originates from a research project aiming to relate two Trinity College London examinations to the CEFR (Papageorgiou 2007a) using the methodology of the Council of Europe's Manual for relating exams to the CEFR . In order to address the research questions, data were collected during the familiarization activities suggested in the Manual. In particular, 12 project participants were asked to categorise 124 CEFR descriptors into levels, during 4 different administrations. These descriptors (30 speaking, 25 writing, 19 listening, 20 reading and 30 global) from Tables 1 and 2 in the CEFR (Council of Europe 2001: 24-27) were provided to the participants without any indication of the level they belong to (from A1 to C2). The participants were then asked to indicate the level of each descriptor and discuss with the

group their reasons for choosing a particular level. The discussions were recorded for clarification of comments and for further data analysis.

The total number of judgements was 5800 as can be seen in Table 1. It should be noted that the participants were involved in a fifth rating session for the reading descriptors only, as part of a different project.

Table 1 Administration of tasks and number of ratings

Descriptors	N	Number of judges per administration					Ratings
		Sept 2005 1st	Sept 2005 2nd	November 2005	February 2006	July 2006	
Speaking	30	12	12	10	11	-	1350
Writing	25	12	12	10	11	-	1125
Listening	19	12	12	10	11	-	855
Reading	20	12	12	10	11	11	1120
Global	30	12	12	10	11	-	1350
Total	**124**						**5800**

3. Data analysis

In order to analyse the rank-ordering of the descriptors by the participants, the many-facet Rasch model (Linacre 1994) operationalised by the computer program FACETS (Linacre 2005) was employed (see McNamara 1996 for a detailed discussion). This program was also used by North (2000) in the project that developed the CEFR scales.

The many-facet Rasch model, primarily designed for tests involving raters (Linacre 1994: 2), thus speaking and writing, was chosen for the study reported in this paper, because it could provide estimates of descriptor rank-ordering by taking into account rater severity and differences across different occasions. Therefore, if a rater is too strict (i.e. assign descriptors at a lower level compared to the other raters) or if raters are stricter in one occasion over another, FACETS will take this into account when producing the descriptor scaling. Three facets of measurement were defined for the analysis:

1. the CEFR descriptors, comprising the item facet for which the participants provided ratings in terms of level
2. the participants, comprising the rater facet
3. the different administrations comprising the occasion facet

FACETS provides a wealth of tables and statistics which offer valuable information as to the rank-ordering of the descriptors. Due to space limitations and to avoid technical jargon, the discussion of results in Section 4 will be primarily based on a

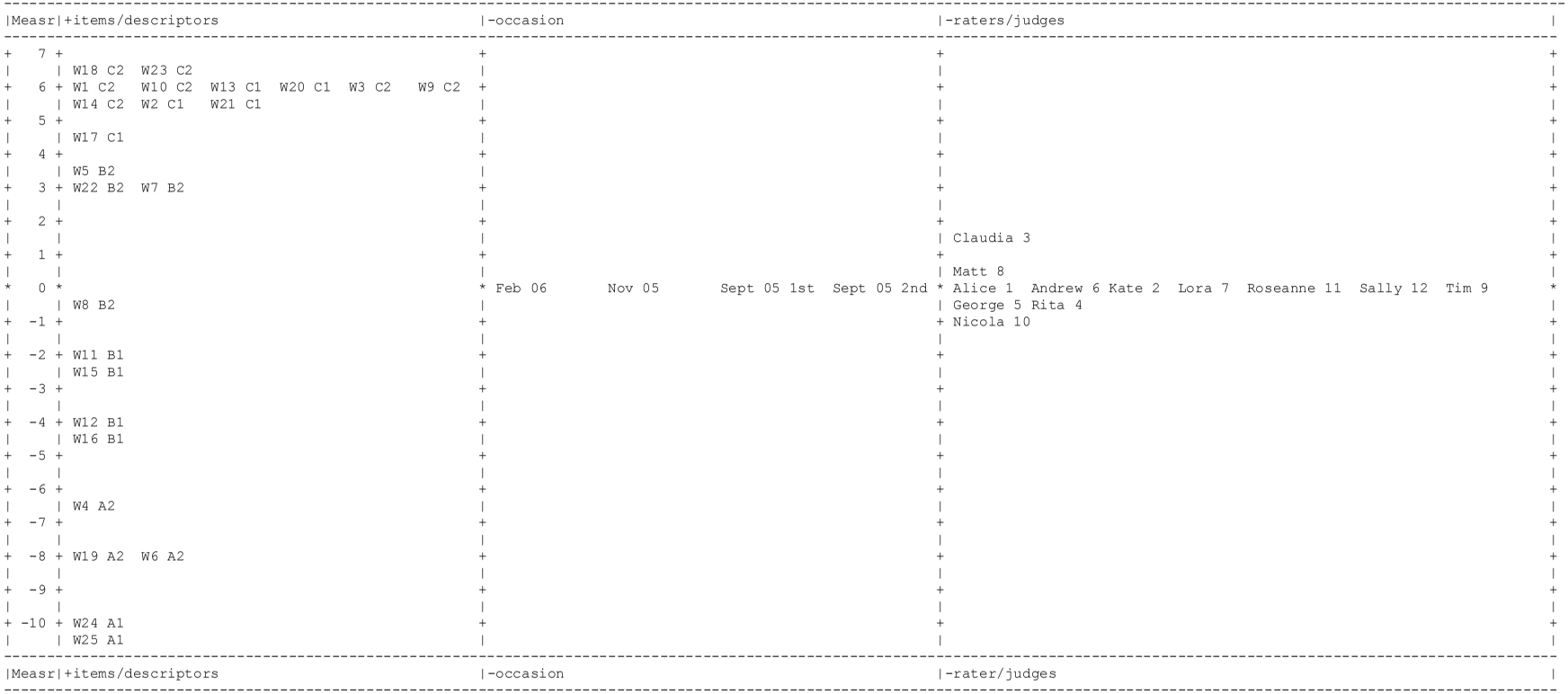
figure called “all-facet vertical summary”. For a more technical discussion see Papageorgiou (2007b).

4. Results

The vertical summary for writing is presented in Figure 1. Similar figures were produced for all five sets of descriptors, but only writing will be discussed here due to space limitations; however it should be noted that results for sets of descriptors were similar.

The first column of the vertical summary (Measr) is the logit scale, an interval scale centered on 0. The second column shows the descriptor IDs (see Appendix for details), which are accompanied by the level in which they appear in the CEFR for ease of reference. For example the lowest descriptors for writing are W24 and W25, which both belong to Level A1. Lower level descriptors appear at the bottom of the scale and higher descriptors at the top. Similarly, the four occasions appear in the third column. Occasions higher on the scale generated stricter ratings, i.e. levels assigned to descriptors were lower than in the other occasions. With regard to raters in the fourth column, the vertical summary shows that a rater higher on the scale is stricter than the others, thus assigning lower levels to the descriptors compared to other raters.

Figure 1 All-facet vertical summary for Writing



The progression from lower to higher levels in Figure 1 appears to be similar to the intended one in the CEFR scales, starting with A1 descriptors at the bottom of the scales and continuing with A2 descriptors and so on. However, this is not the case for the higher levels, in particular C1 and C2 which are not as clearly separated as the lower levels. This pattern was observed with all descriptor sets, suggesting that the higher-level CEFR descriptors did not indicate a clear progression from level to level, so that the raters could order them in three distinct groups (i.e., B2, C1 and C2). Therefore in relation to the first research question (rank ordering of the descriptors from lower to higher levels), it appeared that, despite the generally correct ordering of the descriptors by the raters, higher levels were not clearly understood.

With regard to the second research question (reasons for incorrect rank-ordering), the recordings of the group discussions were examined. A systematic qualitative analysis of reasons for incorrect scaling through verbal protocols or interviews was impossible due to time limitations. Nevertheless, the recordings of the group discussions provided some useful information as to why the raters' scaling for specific descriptors did not agree with the CEFR. Notes were taken from those parts of the recordings where the raters talked about the descriptors revealing the following reasons for incorrect scaling:

1. Lack of definition of particular words. Some words in the descriptors were not clearly defined and this resulted in incorrect rank-ordering as the raters seemed to rely on them when choosing a level. . For example, ''most' and 'without too much' (L19 and L18; see Appendix) probably resulted in the judges' perception of difficulty in the opposite order than the one presented in the CEFR. These two descriptors were found in the same reverse order in Kaftandjieva and Takala (2002: 125), which they also attributed to the use of 'most' and 'without'.
2. Inconsistent use of key words. Some descriptors attracted comments for inconsistent use of key words. For example, 'complex' in the writing descriptors, appears in both C1 and C2 (e.g., W14 and W20; see Appendix) in such a way that it made the distinction between the two levels difficult for the judges.
3. Amount of detail. When a more detailed description of language behaviour appeared in a descriptor, the participants tended to overestimate the level of this descriptor.

Interestingly, the aforementioned reasons for incorrect rank-ordering bear similarities to findings of the other CEFR studies (Alderson *et al.* 2006, Generalitat de Catalunya

2006, Kaftandjieva and Takala 2002). This may signify issues with the descriptors that were not confined to this group of raters, but relate to inadequacy of the descriptors when used in specific contexts. Furthermore, it may suggest that the original CEFR scales need some adjustment as to the level of specific descriptors, which seem to describe lower or higher levels of ability than the intended ones in the CEFR.

Finally, with regard to third research question (effectiveness of training), it is clear from Figure 1 that results across occasions were similar. This might indicate that training was not effective in showing participants the level of those descriptors that they placed at wrong levels, thus wrong placements persisted across time. However, it might also be the case that the participants were in general very good at rank-ordering the items (indicated by the Measr column in Figure 1) and therefore it was not realistic to expect any changes in ratings across different occasions.

5. Conclusion

The findings of this study have an important implication about the claims that examination providers make with regard to links to CEFR: If trained participants have problems distinguishing progression of language ability in the higher CEFR levels, then it is not clear how meaningful it is to state that Examination A is at C1 and Examination B at C2. This is extremely important for the CEFR linking context, as there are a number of well-known language examinations around the world whose providers claim that their scores relate to these levels.

Perhaps future research can address this issue by examining the differences of the higher levels (B2, C1 and C2) both in terms of quantity (i.e. how many things learners can do with the language) and quality (i.e. how well can learners do things with the language). Research on the development of writing ability at the different IELTS bands (Banerjee, Franceschina, and Smith 2007) employed measures from SLA research. This provides some support to Alderson's (2005b, 2007) point that any contribution from SLA can enhance our understanding of how language ability progresses in the CEFR levels. The CEFLING Project[1] at the University of Jyväskylä, Finland, employs SLA research to examine how second language proficiency develops from level to level, which might lead to a better understanding of the differences among the top three CEFR levels.

[1] http://www.jyu.fi/hum/laitokset/solki/en/research/projects/cefling

References

Alderson J.C. (2005a). The challenge of (diagnostic) testing: Do we know what we are measuring? Paper presented at the 27th Language Testing Research Colloquium (LTRC), Ottawa, Canada.

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

Alderson J.C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox and M. Wesche (eds), *Language testing reconsidered: Proceedings of the 27th Language Testing Research Colloquium (LTRC)*. Ottawa: University of Ottawa Press, 21-39.

Alderson J.C., N. Figueras, H. Kuijper, G. Nold, S. Takala, and C. Tardieu (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: 3-30.

Banerjee J., F. Franceschina, and A.M. Smith (2007). Documenting features of written language production typical of different IELTS band score levels. Paper presented at the Fourth Annual Conference of EALTA, Sitges, Spain. http://www.ealta.eu.org/conference/2007/docs/pres_friday/Banerjee%20et%20al.pdf.

Brindley G. (1998). Describing language development? Rating scales and second language acquisition. In L.F. Bachman and A.D. Cohen (eds), *Interfaces between second language acquisition and language testing research* Cambridge: Cambridge University Press, 112-140.

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

Council of Europe (2003). *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Preliminary pilot version.* Strasbourg: Council of Europe.

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

Fulcher G. (2004a). Are Europe's tests being built on an 'unsafe' framework? 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: 253-266.

Generalitat de Catalunya (2006). *Proficiency scales: The Common European Framework of Reference for Languages in the Escoles Oficials d'Idiomes in Catalunya*. Madrid: Cambridge University Press.

Huhta A. and N. Figueras (2004). Using the CEF to promote language learning through diagnostic testing. In K. Morrow (ed.), *Insights from the Common European Framework*. Oxford: Oxford University Press, 65-76.

Kaftandjieva F. and S. Takala (2002). Council of Europe scales of language proficiency: A validation study. In J.C. Alderson (ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies*. Strasbourg: Council of Europe, 106-129.

Keddle J.S. (2004). The CEF and the secondary school syllabus. In K. Morrow (ed.), *Insights from the Common European Framework*. Oxford: Oxford University Press, 43-54.

Linacre J.M. (1994). *Many-facet Rasch measurement* (2nd ed). Chicago: MESA Press.

Linacre J.M. (2005). FACETS Rasch measurement computer program version 3.58. Chicago: Winsteps.com.

McNamara T. (1996). *Measuring second language performance*. Harlow: Longman.

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

North B. (2005). *The CEFR levels and descriptor scales*. Unpublished manuscript, from a paper presented at the 2nd International Conference of ALTE, Berlin, 19-21 May 2005.

North B. and G. Schneider (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15: 217-262.

Papageorgiou S. (2007a). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference: Piloting of the Council of Europe draft manual.* (Final project report). London: Trinity College London. http://www.trinitycollege.co.uk/resource/?id=2261

Papageorgiou S. (2007b). Setting standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference. Unpublished PhD thesis, Lancaster University.

Weir C.J. (2005). Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. *Language Testing* 22: 281-300.

Appendix

Descriptors for Writing and ID numbers	Level
W1 I can write summaries of professional or literary works.	C2
W2 I can write detailed expositions of complex subjects in a letter underlining what I consider to be the salient issues.	C1
W3 I can write reviews of professional or literary works.	C2
W4 I can write a very simple personal letter, for example, thanking someone for something.	A2
W5 I can write clear detailed text on a wide range of subjects related to my interests	B2
W6 I can write short simple messages relating to matters in areas of immediate need.	A2
W7 I can write an essay passing on information or giving reasons in support of or against a particular point of view.	B2
W8 I can write letters highlighting the personal significance of events or experiences.	B2
W9 I can write clear smoothly flowing text in an appropriate style.	C2
W10 I can write complex articles.	C2
W11 I can describe impressions.	B1
W12 I can write personal letters.	B1
W13 I can write detailed expositions of complex subjects in a report underlining what I consider to be the salient issues.	C1
W14 I can write complex letters.	C2
W15 I can describe experiences.	B1
W16 I can write simple connected text on topics which are familiar or of personal interest.	B1
W17 I can express myself in clear well-structured text expressing points of view at some length.	C1
W18 I can write complex reports.	C2
W19 I can write short simple notes relating to matters in areas of immediate need.	A2
W20 I can write detailed expositions of complex subjects in an essay underlining what I consider to be the salient issues.	C1
W21 I can write different kinds of texts in an assured personal style appropriate to the reader in mind.	C1
W22 I can write a report passing on information or giving reasons in support of or against a particular point of view.	B2
W23 I can present a case with an effective logical structure, which helps the recipient to notice and remember significant points	C2
W24 I can write a short simple postcard, for example, sending holiday greetings.	A1
W25 I can fill in forms with personal details, for example, entering my name, nationality and address on a hotel registration form.	A1

Descriptors for Listening and ID numbers	Level
L1 I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.	C2
L2 I can understand most TV news programmes.	B2
L3 I can recognise familiar words and very basic phrases concerning myself when people speak slowly and clearly.	A1
L4 I can recognise familiar words and very basic phrases concerning immediate concrete surroundings when people speak slowly and clearly.	A1
L5 I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when I am spoken to relatively slowly and clearly	B1
L6 I can recognise familiar words and very basic phrases concerning my family when people speak slowly and clearly.	A1
L7 I can understand extended speech provided the topic is reasonably familiar.	B2
L8 I can understand lectures provided the topic is reasonably familiar.	B2
L9 I can understand extended speech even when it is not clearly structured.	C1
L10 I can understand films without too much effort.	C1
L11 I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc.	B1
L12 I can catch the main point in short, clear, simple messages.	A2
L13 I can understand the majority of films in standard dialect.	B2
L14 I can understand extended speech even when relationships are only implied and not signalled explicitly.	C1
L15 I can catch the main point in short, clear, announcements.	A2
L16 I can understand phrases and normal vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local geography, employment).	A2
L17 I can understand follow even complex lines of argument provided the topic is reasonably familiar.	B2
L18 I can understand television programmes without too much effort	C1
L19 I can understand most current affairs programmes.	B2