**Author(s):** Huhta, Ari; Alanen, Riikka; Tarnanen, Mirja; Martin, Maisa; Hirvelä, Tuija

**Title:** Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages

**Year:** 2014

**Version:**

**Please cite the original version:**

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. Language Testing, 31(3), 307-328. https://doi.org/10.1177/0265532214526176

Assessing learners' writing skills in a SLA study - Validating the rating process across tasks, scales and languages


## Abstract

There is relatively little research on how well the CEFR and similar holistic scales work when they are used to rate L2 texts. Using both multifaceted Rasch analyses and qualitative data from rater comments and interviews, the ratings obtained by using a CEFR-based writing scale and the Finnish National Core Curriculum scale for L2 writing were examined to validate the rating process used in the study of the linguistic basis of the CEFR in L2 Finnish and English. More specifically, we explored the quality of the ratings and the rating scales across different tasks and across the two languages. The relationship of task peformance across the scales and languages was also examined. The kinds of analyses reported here are relevant to other SLA studies that use rating scales in their data gathering process.



**Keywords**

validation, rating process, CEFR scales, tasks, L2 writing, L2 learning


## Background

Since the 1990s the number of empirical studies combining language testing and second language acquisition (SLA) research has slowly grown (see, e.g., Bachman & Cohen, 1998). The introduction of the Common European Framework of Reference, CEFR, (Council of Europe, 2001) for languages has created an interest in Europe in the study of the relationship between the communicative L2 development, e.g., functions as described in the CEFR levels, and the development of the linguistic skills, e.g., vocabulary and structures (Bartning, Martin & Vedder, 2010). Can specific linguistic features be associated with specific proficiency levels? To what extent are such associations dependent on the learners' first language or the language being learnt? An interest in such questions has characterized the work of, for example, the European SLATE (Second Language Acquisition and Language Testing in Europe) network of researchers (see www.slate.eu.org).

The CEFR has become an essential element of European language education (see Huhta, 2012), largely thanks to its political backing (Chalhoub-Deville, 2009). This has been somewhat controversial. On the one hand, the CEFR has promoted an action-oriented view of language,

criterion-referenced assessment based on proficiency levels, and the concept of language profiles. It has also raised awareness of the principles of valid and fair assessment. Importantly, the CEFR has provided shared concepts for discussing language use and learning. However, the CEFR has often been implemented in a normative fashion that violates its intended flexible, concertina-like use as a reference tool (North, 2007). The CEFR levels have been applied for educational target-setting (e.g., in curricula) and, as a policy level tool, for setting language requirements for such high-stakes purposes as citizenship. This is not necessarily invalid if the CEFR scale is applied appropriately, transparently and based on empirical evidence, which, however, is not always the case. The very generic nature of the CEFR has also been criticized, particularly the descriptors and whether they can differentiate between different proficiency levels (e.g., Galaczi, 2013).

The power of the CEFR has brought attention to its uncertain compatibility with findings from SLA research or its suitability for young learners (Hulstijn, 2007, 2010; Little, 2007; North, 2007). Although there is evidence about the coherence of the descriptors that form the CEFR scales (e.g., North, 2000; Kaftandjieva & Takala, 2002), the scales have been criticized for ambiguities and inconsistencies (Alderson, 2007, 661; see also the discussion section below). For studies that examine links between linguistic features and the CEFR levels, a particularly important question is whether the straightforward approach of using CEFR scales to place learner performances on proficiency levels really works. Until very recently, evidence about the suitability of the CEFR scales for rating purposes has been lacking but studies by Chen (2009), Kuiken, Vedder and Gilabert (2010), Forsberg and Bartning (2010), Carlsen (2010) and Eckes (2012), each applying slightly different approaches to CEFR-related rating, suggest it may be possible. More generally, rating scales are not commonly used in SLA studies (see Tremblay & Garrison's, (2010) review).

The present study adds to our knowledge about the rating procedures used in SLA research and about the factors that can affect the quality of the ratings. More specifically, we want to know if the unmodified CEFR scale and a local modification of the CEFR scale are suitable for rating.

## Data and methods

### Participants and tasks

Research reported here is based on CEFLING, *The linguistic basis of the Common European Framework levels: Combining second language acquisition and language testing research* (see

www.jyu.fi/cefling and Alanen, Huhta & Tarnanen, 2010; Alanen, Huhta, Jarvis, Martin & Tarnanen, 2012), which was a three-year (2007-09) cross-sectional study on writing in L2.

The 13–16 year old participants of the CEFLING study were 250 Finnish-speaking learners of English as a foreign language and 226 immigrants from different L1 backgrounds studying Finnish as a second language in Finnish-medium comprehensive schools in grades 7-9. The rater comments reported in the qualitative part of this article include alsodata from TOPLING (see www.jyu.fi/topling), a subsequent, longitudinal three-year study. The tasks, scales, raters, and rating procedures in both studies were the same. The age range of the writers in TOPLING was larger: grades 1 - 12 for L2 Finnish, grades 3 - 12 for English, and 7 - 12 for Swedish (this last langauge was not included in CEFLING).

The participants completed four writing tasks. To ensure that the participants were familiar with the tasks, school textbooks and a national language examination were used as a resource in task design. After extensive piloting, five different tasks were used in the actual data collection in both languages:

- Task 1: Informal email message to a friend
- Task 2: Informal email message to the teacher
- Task 3: Formal message to an Internet store
- Task 4: Opinion (response to a given topic)
- Task 5: Story (telling about a personal experience)

The tasks varied stylistically from informal (Task 1) to semi-formal (Task 2) to formal (Task 3). Functionally, the tasks represented various text types (e.g., narrative and argumentative) and functions (e.g., greeting, inquiry, complaint). All learners completed Task 3, 4 and 5 and either Task 1 or 2 (each student was randomly given only one of these two). The informal and functionally less demanding text types were mostly intended for A1 and A2 learners while the others were thought suitable for more advanced learners (see Alanen et al,, 2010, for details).

**Rating scales**

Learner performances were rated by using two scales, a CEFR scale for writing, and the Finnish *National Core Curriculum for Basic Education* (2004) scale for writing. The CEFR scale was created by putting together several genre-specific CEFR writing scales, without modifying their

wording (Appendix 1). We explicitly excluded scales covering the linguistic aspects of performance, to focus the raters' attention to the communicative aspects of the texts. The second scale used in the study, the Finnish National Core Curriculum (NCC) scale is an overall scale for writing, although it, too, contains descriptions of several dimensions of writing (Appendix 2).

The CEFR scales were not originally developed for rating learners' performances (except for one self-assessment grid/scale), although they have later been used for that purpose, too. The issue with the CEFR scales is that they lack features typical of scales designed specifically for assessment purposes such as references to errors in learners' performance. Most CEFR scales focus on defining elements such as tasks, activities and texts, which are features typical of scales designed for descriptive and reporting purposes (see Alderson, 1991, on different scales). Thus, the validity of the CEFR scales for rating speaking and writing cannot be automatically assumed, although some previous studies are promising in this regard (e.g., Chen, 2009; Carlsen, 2010). One approach to tackling this problem is to modify CEFR scales to make them more rater-friendly, as, for example, Harsch and Martin(2012) did in their study. In any case, the suitability for rating of any scale needs to be examined before the ratings obtained with it can be trusted.

The NCC scale developed in Finland in the early 2000s differs from the CEFR scales although much of its content comes from the CEFR (Appendix 2). The scale is used in the mainstream education in Finland, from primary to upper secondary levels, to describe learning, teaching and assessment targets for foreign and second languages. To provide learners with more easily reachable learning targets the original CEFR scale levels were divided into two or three sub-levels. References to limitations and errors in learners' performances were also added to help assessments (Hildén & Takala, 2007; Tarnanen & Huhta, 2008). Despite the importance of this scale, its suitability for assessment purposes has not been examined (Tarnanen & Huhta, 2008). An obvious issue with the NCC is, for example, that is comprises ten levels, which may be too many to distinguish in rating.

If a more fine-tuned scale such as the NCC can be used successfully, it may allow us to obtain more detailed information about how linguistic features relate to proficiency levels than the 6-point CEFR scale does. If we discover, for example, that a particular linguistic feature occurs significantly more often at level B1 than at A2, applying a scale like the NCC for rating might reveal the more precise point at which the change happens , i.e., whether it occurs at the lower or higher end of the particular CEFR level.

**Raters and the rating procedure**

A total of eight raters assessed the English performances and eleven raters assessed the Finnish performances. All raters were language education professionals in the language concerned: language teachers, lecturers or professors, some of them also experienced raters. A linked rating design ensured that each Finnish L2 performance was rated by three and each English L2 performance by four different raters (the difference in the numbers was due to the availability of raters).

The raters completed a two-stage training process. First both Finnish and English raters familiarized themselves with both rating scales (NCC and CEFR) and benchmark performances obtained from a national language examination that uses a CEFR level referenced rating system. The English raters also studied the five writing samples representing levels A2 - C2 available for English at the Council of Europe website (no samples for A1 in English were available from the Council . After familiarisation the raters assessed Finnish or English sample performances obtained during the piloting of the writing tasks. These were then discussed, and the samples that were most unanimously rated were selected to serve as the benchmarks in the actual CEFLING study. The raters first assessed all the performances assigned to them with the CEFR scale and after about a month they re-rated the texts with the more fine-grained NCC scale. On both occasions, each rater gave the texts only one overall holistic CEFR or NCC rating.

**Data**

Our data were of two types. First, we had the ratings for the five tasks in two languages, obtained by using two different rating scales (CEFR and NCC), as was described above. These were analysed with Facets, the multifaceted Rasch measurement software (Linacre, 2009).

The second type of data was qualitative and was based on rater interviews and raters' written comments. A semi-structured interview with six raters was conducted after the completion of the ratings to examine their perceptions of the two scales, in particular, but also of the tasks and the entire rating process. For additional rater perceptions of the scales and other aspects of the rating process, the approximately 4300 comments written by the raters on the rating forms during the CEFLING and TOPLING studies were collected. The interviews and comments were analysed and categorized qualitatively.

**Research questions**

RQ1: How do the chosen rating procedures work across the different tasks and both languages?

- RQ1a: What is the quality (consistency) of the ratings?
- RQ1b: Do both scales work as rating scales (e.g., can the levels be distinguished from each other) in both languages?
- RQ1c: How comparable are the two scales?
- RQ1d: Do (some of) the raters use them differently?
- RQ1e: Do the raters use other criteria than those mentioned in the scale descriptors?

RQ2: How do the tasks used in this study function across the scales and languages?

- RQ2a: Does task performance vary across different task types?
- RQ2b: Do the ratings vary depending on the scale used or the language rated?
- RQ2c: Does the task systematically affect (some) raters' ratings?

## Results

### Quality of ratings

The quality of the ratings refers here to raters' consistency (does the rater maintain the same level of severity/leniency across all learners?) and their comparability (do the raters differ in terms of leniency/severity?). The ratings were analysed with the multifaceted Rasch programme Facets following the guidelines of the Facets manual (Linacre, 2009) and previous studies (e.g., Lunz, Wright & Linacre, 1990; McNamara, 1996). The examination of the Infit indices revealed that one of the raters of English was too inconsistent when using the NCC scale (the Infit value was 1.6, while the acceptable range is usually 0.5 - 1.5). Since every text was rated by four raters, it was possible to remove the misfitting rater from the analyses. In addition, 20-40 individual data points (ratings) were removed from both the Finnish and English datasets because they were statistically misfitting. These misfitting ratings came from several different raters, although such aberrant ratings appeared to be more common with some raters than others, even if these raters had not turned out to be too misfitting in the overall rater infit analyses. There were more

misfitting ratings when the NCC scale was used, possibly because of the greater length of that scale compared with the CEFR scale.

Raters' tendency to differ from each other in how severely or leniently they judge performances can also be an issue, and previous research shows how difficult it is to remove such differences (Lunz, Wright & Linacre, 1990; Knoch, 2010). As reported above, all our raters except one 'fitted the model' in the sense that they rated consistentently enough to be trusted as raters. However, it is clear that they were not equally severe, as Appendix 3 and 4 show. The statistical tests included in the Facets for the differences between different elements (e.g., individual raters) within a particular facet (e.g., raters) indicated that, overall, raters' severity differed significantly. For example, in the statistical sense the nine raters of English using the CEFR scale represented 6-7 severity levels, i.e., almost all of them differed from the others in terms of severity. The results were quite similar for the raters of Finnish and also when the raters used the NCC rating scale instead of the CEFR.

Obviously, large variation between raters is a problem and should be addressed in some way. In our case, the CEFR or NCC levels awarded to the texts should not change even if one of the raters happened to be considerably more/less severe than the others.

Perhaps the best way to decrease the effect of the inevitable variation in rater severity is the use of multiple ratings, because, first, the effect of a very severe/lenient rater is diminished by the other, more moderate raters, and, second, it is possible to remove raters from the analyses if they differ too much. The Facets programme adjusts its results on the basis of the analysis of the raters' relative severity by lowering the scores given by a lenient rater and raising those given by a severe rater. However, such adjustment is probably limited and a very severe/lenient rater may bias the results. We have not come across with discussion of this issue in the literature on Facets. The statistical test of rater homegeneity in the Facets (the Separation index) is of limited use as almost any sizable group of raters is likely to differ significantly in terms of severity. Solutions may be sought by examining how raters align with the rating scale levels to see what practical consequencies differences in raters' behaviour have, as is done below.

Appendices 3 and 4 display how the analysed facets align with each other. From left to right, the figures show: (1) the logit measure scale created by Facets (this is the common yardstick, a true interval scale, agaist which the facets align), (2) the learners ('examinees'): the more able students are on the top and the less able at the bottom, (3) the raters (lenient raters up, severe raters down), (4) the tasks (difficult tasks up, easier tasks down), (5) the CEFR rating scale, and (6) the NCC rating scale. They show that the difference in rater severity in our study equals half a

scale point on the 6-point CEFR scale in both languages. In the 10-point NCC scale, the most lenient and the most severe rater of English are separated by 1½ levels; for Finnish the distance is 2½ levels. Appendix 4 shows that Rater 11 differs considerably from the other raters of Finnish; without this rater, all raters of Finnish would fall within half a CEFR band or within one NCC band (it was checked if removing Rater 11 significantly changed the results reported in this article: it did not). For English, no rater stands out, so there appears to be no reason to consider removing raters simply because of their severity or leniency.

The above analyses suggest that the quality (consistency and comparability) of the ratings of the texts collected in the CEFLING project was high enough so that the placements of the scripts on the CEFR and NCC levels can be trusted (see RQ1a above).

**Quality and comparability of the rating scales**

For research into the applicability of the two scales for rating purposes, it was necessary to examine their quality. The main question (RQ1b) is: Can the raters distinguish the scale levels from one another? It is also of interest to see if the main levels of the two scales align with each other the way they are supposed to (RQ1c). That is, the major scale boundaries of the two scales should match closely because the NCC scale was designed by adding new content to, and by splitting, the broad CEFR levels.

Appendix 5 shows the results of the Facets analysis of the CEFR and NCC scales based on the English and Finnish ratings. The Appendix displays the number and percentage of the texts rated at each level with the two scales and in the two languages. The important information is the Rasch-Andrich Threshold measures which show where in the logit scale each level boundary is placed (the same logit scale can be found in the first column in Appendices 3 and 4). The threshold values should increase in a linear fashion and the thresholds should be separated well enough before we can argue that the raters, as a group, were able to distinguish all the levels of the scale. It is clear that all threshold values increase from the bottom of the scale to the top. As to the width of each level, Linacre (2002) argues that the minimum distance between thresholds is related to scale length: for instance, for a 5-point scale, 1.0 logit separation suffices. According to this kind of calculation, both the CERF and NCC scale levels seem wide enough to be meaningfully distinguishable for rating purposes.

The content of the CEFR and NCC scales differs, as Appendices 1 and 2 show, but the NCC scale is intended to have the same basic vertical structure as the CEFR scale (RQ1c). Thus, the

same amount of proficiency should be required to reach, say, A2 on the CEFR and A2.1 on the NCC. Facets analyses show that the main scale level boundaries are quite close, especially in the Finnish dataset (see Appendices 3 and 4). However, the English NCC boundaries were consistently somewhat higher than the corresponding CEFR boundaries, except at the lowest end of the scale. This means that a borderline A1/A2 learner of English was likely to be rated at A2 on the CEFR scale but at A1.3 on the NCC scale.

**Analysis of task performance**

Before learner performances were rated, 24 language experts (12 in each language, including the raters who eventually rated the texts) judged the tasks against the CEFR levels. According to their estimates, Task 4 (opinion) was regarded as most suitable for B2, while Task 1 (message to a friend) was considered the easiest for both languages (A2). Tasks 2 and 5 were considered best suited for B1. The biggest difference between languages was for Task 3: it was considered somewhat more difficult for English than Finnish learners (B1+ vs B1).

**Task performance and the CEFR scale.** Facets analyses were also used to examine the task performance across the two scales and two languages (see RQ2 and RQ2a above). Appendix 6 shows the task measurement report, including the fair-mean (adjusted for rater severity/leniency) averages of the performances in both languages. As the language experts expected, task 4 (opinion) turned out to be the most challenging in both languages. The English learners had more trouble with Task 2 (email to teacher), than was predicted while for the Finnish learners it was the easiest. Task 1 (email to a friend) was more difficult in English than Task 3, an email to an internet store. Also Task 2 (message to teacher) was more challenging than expected.

Although the differences in fair-mean averages between the task performances on the CEFR scale were not great (ranging from 2.01 for Task 3 to 1.87 for Task 4), the separation index of 2.74 obtained for these ratings indicates that the performances could still be separated fairly reliably into roughly three groups, with Tasks 1, 2 and 5 being rather similar but different from Task 3 and 4.

The tasks written in Finnish were also rather similar in task performance; all were slightly over 2.0 (i.e., A2). However, almost all of them were statistically different from each other as the high 4.65 separation index indicates. Only Tasks 1 and 2 (or perhaps 4 & 5) were not clearly separated from each other.

**Task performance and the NCC scale.** Similar results were obtained when the raters used the NCC scales to judge learner performances (the second table in Appendix 6). All tasks fell between fair-mean average of 4 (i.e., A2.1) and 5 (A2.2), with the Finnish results again being slightly higher than English.

The separation of the tasks into different difficulty levels found for the CEFR scale was reversed when the NCC scale was used. Now the English tasks could be separated into over four (4.29) groups, with only Tasks 2 and 4 being indistinguishable in terms of writing ability required by them (Table 3). The Finnish tasks, for their part, could be separated into only 2.34 groups: Tasks 1, 2 and 3 were almost equally challenging (Table 4). Tasks 4 and 5 appeared somewhat more clearly separable from each other in NCC rating than in CEFR rating.

The ranking of the tasks in terms of task performance was also slightly changed when the NCC scale was used. For English, Task 5 (narrative) become more 'difficult' than Task 1 (email to a friend), when they had been indistinguishable in CEFR scale ratings. Otherwise, the ordering of the English tasks remained the same. For Finnish, the poorer separation power of the NCC scale means that it is not possible to say if the ordering of the tasks was changed from the CEFR ratings: all message type of tasks appear equally demanding but easier than Tasks 4 and 5.

In absolute terms, there was, thus, rather little cross-task variation in the students' performance regardless of the language or scale. Learners of Finnish as L2 seemed to slightly outperform the learners of English as FL on all tasks.

**Results of bias analyses**

To further examine the ratings, bias analyses were run with Facets to establish whether there was any significant interaction between (1) raters and tasks (RQ1c), (2) scales and tasks (RQ2b), and (3) raters and scales (RQ1d). Since the number of possible interactions between these facets is quite large only a summary of the findings is given below.

The analyses showed no significant interaction between the scales and tasks: Both scales were applied in the same way across all five tasks. In contrast, a number of slight but significant rater by task interactions were found; the majority of raters in both languages demonstrated some bias for or against one or more of the tasks. No clear patterns could be detected, however: raters' 'preferences' seemed quite idiosyncratic. Finally, rater by scale bias analyses showed that seven of the twelve raters of Finnish L2 performances used one of the two scales more leniently than the other. Three of the twelve English raters demonstrated similar behaviour. Again, no clear

pattern could be detected, as some raters were more lenient when using the CEFR scale while others were more lenient with the NCC scale.

**Qualitative analyses**

To complement the quantitative analyses of the scales and tasks rater interviews were used. The voice of the raters is also heard in the comments they wrote in the rating forms. These data shed light on all the sub-questions of Research Question 1 but in particular on RQ1e that concerns the criteria that the raters actually used when rating performances.

**Interviews of raters.** A semi-structured interview with three raters of English and three raters of Finnish raters was conducted after completion of the ratings to examine their perceptions of the two scales, in particular, but also of the tasks and the whole rating process.

Overall, the raters found the CEFR scale easier to use than the NCC scale. The CEFR scale was also considered to be more positive in orientation, more "can do centred", whereas the NCC scale was seen to focus more on mistakes and limitations in proficiency. Raters mentioned such NCC linguistic descriptors as *can write simple words and structures accurately, but makes mistakes in less common structures and forms* (level A2.2). This, they felt, made them concentrate more on the weaknesses in performances. Consequently, the NCC scale was considered to be more severe. On the other hand, the NCC *more concrete* and *detailed* for the raters, whereas the CEFR scale was *vague* and *ambiguous*.

Raters' opinions on the number of levels in the scales were contradictory. One rater thought the CEFR scale was easier to use because "[it] was not divided into so many levels as the NCC scale" but another preferred the NCC scale as "even A1 is divided into three levels and there are things that you can hold on to and try to use them as criteria". The great number of levels on the national scale was also considered a problem, however.

On the whole, the descriptions of the CEFR levels A2 and B2, and the NCC sub-levels for A1 and B1 were thought to be helpful. The following points in the CEFR scale were singled out as particularly useful: *simple phrases and sentences linked with simple connectors* (A1), *can write everyday aspects of his/her environment* (A2) and *marking the relationship between ideas in clear connected text and following established conventions of the genre concerned* (B2). Positive elements in the NCC scale included *cannot express him/herself freely* (A1.1), *can manage to write in the most familiar, easily predictable situations related to everyday needs* (A1.3), *can provide some supporting detail to the main ideas and keep the reader in mind* (B1.2).

The suitability of the scales for assessing young learners worried some raters. They also discussed the view of language and language ideology underlying the scales and their compatibility with the learning process. The raters found several descriptors unsuitable for assessing *second* language learners' (i.e., learners of Finnish) performances:

> for example at A2 level [it says] that there are very short simple basic descriptions of events and so on so again I think this doesn't work very well … with migrant students as they can write terribly long texts and it can be terribly idiomatic in many ways but at the same time it can also be awfully inaccurate with basic [errors]

Of the sublevels of the NCC scale, A2.1 and A2.2 were singled out by the interviewees as particularly difficult to tell apart:

> descriptors for A2.1 and A2.2 are fairly similar so is the difference between them in the number of errors

> the only difference between these levels is in the criterion of structures so it guides you to make a decision based on structures

The raters reported using the scale descriptors selectively for different tasks; i.e., they chose the descriptors that seemed most relevant to the task to be rated. They suggested the scales could be improved by adding references to *task completion*, *comprehensibility of the text*, *register* and more generally to *pragmatic elements*.

> I expected that comprehensibility would somehow be referred to at the A level as some kind of a criterion but it was missing

Although the raters found the genre-related approach of the CEFR scale useful some thought it should contain more text types relevant for secondary school pupils such as essays, reports and examination responses. Also, the ability to express opinions was seen missing at the A levels although even beginners were clearly able to perform this language function.

**Raters' comments in rating forms.** For additional rater perceptions of the scales and the rating process the approximately 4300 comments written by the raters on the rating forms during the

CEFLING and a subsequent, longitudinal study were collected, classified, and analyzed. The longitudinal part of the study covered students from primary to university level.

While the great majority of ratings were not commented, some texts elicited comments from several raters. Thetexts by university and primary level students collected more comments than those of writers at the secondary level. As nearly all comments referred to problems in rating, a low number of comments may indicate a lack of problems. Based on this kind of analysis, the two scales studied might fit best the writing of the 16–19 year old writers in the upper secondary schools (gymnasia), as they were the least often commented texts.

Very short texts elicited comments across all age groups, especially when the rater was considering whether the text was at A1 or below A1. Such texts were found difficult to rate due to a lack of evidence. Also off-topic or humorous texts with maybe questionable content were found problematic. Code-switching (use of languages other than the target language) and problems of comprehension (e.g., handwriting) were frequently mentioned. Interestingly, text content in relation to the task and the quality of argumentation were rarely mentioned, even at the upper levels where argumentation is present in the scales.

The most frequently mentioned issue for the youngest writers was task completion whereas for the older students they were textuality and cohesion. As in the rater interviews, comprehensibility was frequently commented on: It is only mentioned at B1 in both scales but many raters wished it were a criterion also at the lower levels. A similar finding surfaced also in the study by Toropainen, Härmälä & Lahtinen (2012) which compared rater comments in Swedish language data written by L2 learners in Finland and by a similar group of native speakers of Swedish in Sweden.

The rating scales do not explicitly refer to the length of the text or to task completion. Nevertheless, both our raters and those in the study of Toropainen et al. (2012) mentioned them frequently. Clarity of sentence boundaries at the A levels also came up as a problem for rating Finnish L2 performances. These issues suggest changes in the NCC scale might be needed in the future: if the raters commonly use criteria that are not mentioned in the scales, maybe they should be included?

In addition to the issues discussed above, the concept of *simple* raised many comments. *Simple* to some raters seemed to equal short. Others connected it with limited content, which could be another way of addressing the issue of task completion. Yet others used the word to describe syntax or vocabulary. The problem clearly lies in the scales. In the CEFR scale (Appendix 1), the word *simple* is used very frequently. At A1 level simple occurs in the descriptors three times, at

A2 13 times (!), and at B1 four times. B2 descriptors contain neither the word *simple* nor its opposite *complex*. At C levels *complex* appears once at C1 and once at C2. At the levels A2 and B1 there are also some other words related to simple such as *basic* at A2 and *straightforward* at B1.

As much of our data is at the A2 level, no wonder the abundance of the word *simple* elicited comments. In the CEFR scales *simple* and its counterparts seem to be connected to all the issues raised in the comments. It relates to short (simple phrase, simple note/letter/message). It seems to refer to content (simple postcard, simple biography). It can also be read as a syntactic term: a simple sentence in the grammatical parlance of many language professionals refers to a sentence which consists of one clause, as opposed to compound and complex sentences. Also a simple phrase could refer to a grammatically simple noun or verb phrase, like a bare noun or a verb with only one dependent constituent. And what about the internal structure of a clause? Is a sentence simple if it consists of one clause but the clause contains noun and verb phrases that are complex? Such contemplations of the meaning of *simple* can lead raters to very different interpretations (for more details, see Martin, 2013).

## Discussion

This study set out to validate two scales which were originally intended to describe the development of L2 writing skills but are also applied for rating purposes. As the rating process involves tasks, languages and rater behaviour, these aspects are a part of the study. The validation of these is an important phase in a study aiming at following L2 linguistic development across the communicative levels determined by the scales.

A major finding was that the CEFR writing scales functioned adequately for rating purposes across all five writing tasks in both languages despite the fact that their rather general, descriptive nature makes them not ideal rating scales. This result supports the findings of the few existing studies that have used unmodified CEFR scales for rating (Chen, 2009; Carlsen, 2010; Forsberg & Bartning, 2010) but the current study provides a more detailed analysis of the qualities of the scale and the raters who used it. Facets analyses indicated that trained raters, as a group, can work consistently with these scales and distinguish the CEFR levels. Evidence from rater interviews was mixed: some raters felt the CEFR scale descriptions were too vague but others liked the fact that the scale has only six levels. These perceived issues with the scale did not, however, prevent the raters from using it consistently.

Also the Finnish NCC scale turned out adequate for rating purposes. The scale level thresholds increased consistently and the levels appeared wide enough to be separable. This was encouraging, as it enables the CEFLING project to examine linguistic development across levels in more detail than by using the 6-point CEFR scale.

As is always the case, the raters differed somewhat from each other in severity and consistency. This rater variation can be a sign of difficulty in working with the scales, at least for some raters. However, scale descriptors are always open to many interpretations and the quality of the rating process never depends on the scale only. Good benchmark performances and training are also essential. Although we believe to have achieved satisfactory quality in rating L2 texts for SLA research purposes in our context, we do not know how representative our view of the meaning of the CEFR levels in English and Finnish is, as it has not been possible to compare our assessments with those by other groups of raters in other countries. Although the CEFR scale descriptors are the same across Europe, the all-important international writing benchmarks are lacking for these two languages (the five English samples available from the Council of Europe do not constitute a representative, internationally validated set of benchmarks).

The writing tasks used in the study were designed to cover a range of proficiency levels; language specialists estimated the tasks represent CEFR levels A2 - B2. However, the Facets analyses placed all tasks within only one CEFR level, even within one NCC level (see Appendix 3 & 4), at high A2 in both languages. Thus, although these tasks vary considerably in terms of, e.g., their genre and level of formality, the learners' performance on all tasks was rather similar. This implies that our intention to include tasks that allow learners with a wide range of proficiencies to demonstrate their writing skills was successful. Even the beginners could, it appeared, show their (limited) proficiency in the more demanding tasks (Tasks 3-5).

There were some differences between the two languages. In Finnish, task performance was in line with the expected task difficulty: the two message tasks which were predicted to be easier turned out that way in the analyses. For English, the message to the teacher (Task 2) turned out to be more demanding than the message to an Internet store (Task 3). Possibly the difficulties in deciding the proper formality level for the message to the teacher made that task harder. Also, while the L2 learners of Finnish found it easy to ask questions about school activities, FL English learners were not used to formulating these questions in English, particularly in the past tense. The real life experience with shopping on the Internet probably contributed to Task 3 being easy; it also allowed some learners to display knowledge of quite a sophisticated knowledge of computer terminology, which may have impressed the raters. Learners' ability to demonstrate functional proficiency in these tasks also suggests that they do similar tasks in their free time.

Although learners' performance was rather similar across tasks, the more detailed statistical analyses indicated that the tasks could be divided into 2-4 separate groups in terms of how demanding they were, depending on the scale and language. The NCC scale, which covers both communicative and linguistic aspects of performance, worked better at separating English task performances from one another than it did for the Finnish task performances. Conversely, the CEFR scale, which focused on communicative adequacy alone (Pallotti, 2009: Alanen et al., 2010), was better in distinguishing the Finnish task performances. The NCC scale also includes a greater number of linguistic criteria (e.g., accuracy, cohesion) in the descriptors. This may have been another factor in the learners' performance on Task 2 in English vs. Finnish.

The reason why, for English, Task 5 (narrative) became clearly more challenging than Task 1 (message to a friend) when the NCC scale was used instead of the CEFR scale may also be related to the fact that the NCC scale includes references to errors and limitations. Task 5 performances were regularly longer than Task 1 performances, and thus, problems in learners' linguistic competence are more evident in Task 5 performances, which may have lowered their NCC ratings.

There was no interaction between the scales and tasks indicating that both scales worked equally well with all tasks. There was, however, considerable interaction between raters and tasks, and also between raters and scales (especially among the raters of Finnish as L2) which was however idiosyncratic, and, thus, did not appear to have any systematic effect on the placement of learners on proficiency levels. The finding highlights the importance of having multiple ratings of the learners' performances, as an individual rater's personal approach to rating a particular task (or scale) may bias the results.

Theraters' oral interviews and written comments did not produce unambiguous findings about differences between the languages or the tasks. They did provide complementary information about the rating scales that partly confirmed the results of the statistical analyses but partly contradicted them. The interviewed raters felt that the NCC scale was more demanding than the CEFR scale, and the Facets analyses concurred, at least for English. On the other hand, some of the scale levels mentioned by the raters as being particularly clear (e.g., A1.1 and A1.3) did not appear so in the statistical analyses. A possible reason for these discrepancies may be that the interviewed raters may not represent the majority opinion. Also, it is possible that even if some raters perceive a particular descriptor or level unclear this does not significantly affect their ratings in practice.

The raters' comments also revealed that they make inferences from the criteria described in the scale and also use criteria not mentioned there. These included text length, task completion, errors (in the case of the CEFR scale), comprehensibility, idiomaticity, and argumentation; the last one does not appear in the lowest level descriptors but it is nevertheless present in many A-level texts. This is in line with Barkaoui's (2010) findings that scale-external criteria such as text length play an important role in holistic ratings. Our raters also felt the weighing of different criteria (e.g., a long, coherent text but with lots of basic errors) problematic. Furthermore, one particular term, *simple*, appeared to cause interpretation problems as it can refer to a range of different features of the text, which were not explicated in the scale descriptors.

## Conclusion

The article reports on the study into several thousand ratings of L2 writing performances in two languages (L2 Finnish and FL English), carried out in the CEFLING project in Finland in 2007-09. The findings shed light on the usability of the scales and tasks used in the study for collecting L2 performances for subsequent SLA analyses. The quantitative analysis of the ratings, scales and tasks showed that the overall rating procedure was reliable enough and the scales valid for the purposes of this study. There were minor differences between languages as to the task difficulty. Thus, the ratings produced a communicatively ascending data set which can be used for studying the linguistic features which appear at the various proficiency levels.

The raters' oral and written comments provided unique information that statistical analyses cannot reveal. The most important findings indicate the potentially multiple interpretations of some expressions used in descriptors, the fact that the ability to express opinions is missing at the A levels although even beginners are able to perform this language function, and that the raters commonly use criteria that are not mentioned in the scales. These results could be used as a basis for further studies with other languages and other data sets to produce evidence for improving the CEFR and NCC descriptors in the future.

## References

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In Bartning, I., Martin, M. & Vedder I. (eds.) Communicative proficiency and linguistic development: intersections between SLA and language testing research (pp. 21-56). *EUROSLA Monograph Series*, 1.
http://eurosla.org/monographs/EM01/EM01home.html

Alanen, R., Huhta, A., Jarvis, S., Martin, M. & Tarnanen, M. (2012). Issues and challenges in combining SLA research and language testing. In D. Tsagari & I. Csepes (Eds.), *Collaboration in language testing and assessment* (pp.15-30). Language Testing and Evaluation Series, Grotjahn, R. &. G. Sigott (general eds). Frankfurt am Main: Peter Lang.

Alderson, J.C. (2007). The CEFR and the need for more research. *Modern Language Journal 91*, 659-663.
Bachman, L. & Cohen, A. (1998). *Interfaces between second language learning and language testing research.* Cambridge University Press.
Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing 27*, 515-535.
Bartning, I., Martin, M. & Vedder, I. (Eds.) (2010) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research.* EUROSLA Monograph series 1. European Second Language Association. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html
Carlsen, C. (2010). Discourse connectives across CEFR-levels: A corpus based study. In I. Bartning, M. Martin & I. Vedder (Eds.) (pp.191-209)
Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. W. Lissitz, editor, *The concept of validity: Revisions, new directions and applications*. Charlotte, NC: Information Age Publishing.
Chen, Y-H. (2009). *Investigating lexical bundles across learner writing development.* Unpublished PhD thesis. Lancaster University.
Eckes, T. (2012). Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling*, 54, 3, 257-283.
Forsberg, F. & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels? A pilot study of written L2 French. In I. Bartning, M. Martin & I. Vedder (Eds.) (pp.133-157)
Galaczi, E. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics* 2013, 1-23 (advanced access: doi:10.1093/applin/amt017).
Harsch, C. & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing 17,* 228-250.
Hildén, R. & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A.Koskensalo, J.Smeds, P.Kaikkonen & V.Kohonen (Eds.) *Foreign languages and multicultural perspectives in the European context* (pp. 291-300). Reihe: Dichtung – Wahrheit – Sprache Bd. 7. Münster: LIT Verlag.
Huhta, A. 2012. Common European Framework of Reference. In Chapelle, Carol (general editor) *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.
Hulstijn, J.H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663-667.
Hulstijn, J.H. (2010). Linking L2 proficiency to L2 acquisition: Opportunities and challenges of profiling research. In I. Bartning, M. Martin & I. Vedder (Eds.) (pp. 233–238).
Kaftandjieva, F. & Takala, S. (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*. (pp.106-129) Strasbourg: Council of Europe.
Knoch, U. (2010). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing 28*, 179–200.
Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin & I. Vedder (Eds.) (pp.81-99)
Linacre, M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement 3*, 2002, 85-106.
Linacre, M. (2009). *A user's guide to FACETS v 3.66.0*. Chicago: Winsteps.
Little, D. (2007). The Common European framework of reference for languages: Perspectives on the making of supranational language education policy. *Modern Language Journal 91*, 645-655.
Lunz, M., Wright, B., & Linacre, M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education 3*, 331-345.

Martin, M. (2013). The complex *simple* – a problematic adjective in the CEFR writing scales. *Nordand – Nordisk t idsskri f t for andrespråksforskning* 8, 2, 63–85.
McNamara, T. (1996). *Measuring second language performance*. Boston: Addison-Wesley Longman.
North, B. (2000). *The development of a common scale of language proficiency*. New York: Peter Lang.
North, B. (2007). The CEFR illustrative descriptor scales. *Modern Language Journal 91*, 656-659.
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics 30*, 590–601.
Tarnanen, M., & Huhta, A. (2008). Interaction of Language Policy and Assessment in Finland. *Current Issues in Language Planning* 9, 3. 262-281.
Tremblay, A., & Garrison, M. D. (2010). Cloze tests: A tool for proficiency assessment in research on L2 French. In M. T. Prior, Y. Watanabe & S.-K. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum. Exploring SLA perspectives, positions, and practices* (pp. 73–88). Somerville, MA: Cascadilla Proceedings Project.

The CEFLING project was supported by a research grant from Academy of Finland.

We would like to thank the following students at the Department of Languages, U. of Jyväskylä, for their contribution in analysing the raters' written comments: Johanna Eloranta. Milla Filppula, Marjaana Göös, Raisa Haikala, Heidi Henttinen, Ulla Huhtala, Janica Häggman, Mari Karppinen, Milja Koski-Lammi, Auli Kotimäki, Maiju Partanen, and Elisa Räsänen.

APPENDIX 1

The collection of CEFR writing scales used for rating purposes in the study

(The entire scale can be found in http://www.jyu.fi/topling)

| | **OVERALL WRITTEN PRODUCTION** | **WRITTEN INTERACTION** | **CORRESPONDENCE & NOTES, MESSAGES, FORMS** | **CREATIVE WRITING & THEMATIC DEVELOPMENT** |
|---|---|---|---|---|
| **A1** | Can write simple isolated phrases and sentences. | Can ask for or pass on personal details in written form. | Can write a short simple postcard.<br>Can write numbers and dates, own name, nationality, address, age, date of birth or arrival in the country, etc. such as on a hotel registration form. | Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do. |
| **A2** | Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'. | Can write short, simple formulaic notes relating to matters in areas of immediate need. | Can write very simple personal letters expressing thanks and apology.<br>Can take a short, simple message provided he/she can ask for repetition and reformulation.<br>Can write short, simple notes and messages relating to matters in areas of immediate need. | Can write about everyday aspects of his/her environment, e.g. people, places, a job or study experience in linked sentences.<br>Can write very short, basic descriptions of events, past activities and personal experiences.<br>Can write a series of simple phrases and sentences about their family, living conditions, educational background, present or most recent job.<br>Can write short, simple imaginary biographies and simple poems about people.<br>Can tell a story or describe something in a simple list of points. |

APPENDIX 2

The Finnish National Curriculum scale for writing (levels A1.1 - A1.3)

(The entire scale can be downloaded from:

http://oph.fi/download/47674_core_curricula_basic_education_5.pdf)

| **Level A1** | | **Elementary proficiency** |
|---|---|---|
| A1.1 | First stage of elementary proficiency | • Can communicate immediate needs using very brief expressions.<br>• Can write the language's alphabets and numbers in letters, write down his/her basic personal details and write some familiar words and phrases.<br>• Can use a number of isolated words and phrases.<br>• Cannot express him/herself freely, but can write a few words and expressions accurately. |
| A1.2 | Developing elementary proficiency | • Can communicate immediate needs in brief sentences<br>• Can write a few sentences and phrases about him/herself and his/ her immediate circle (such as answers to questions or notes).<br>• Can use some basic words and phrases and write very simple main clauses<br>• Memorised phrases may be written accurately, but prone to a very wide variety of errors even in the most elementary free writing, |
| A1.3 | Functional elementary proficiency | • Can manage to write in the most familiar, easily predictable situations related to everyday needs and experiences.<br>• Can write simple messages (postcards, personal details, simple dictation).<br>• Can use the most common words and expressions related to personal life or concrete needs. Can write a few sentences consisting of single clauses.<br>• Prone to a variety of errors even in elementary free writing. |

APPENDIX 3

English CEFLING ratings - CEFR and NCC scales Table 6.0 from Facets output

| Measr | +Examinee | +Rater | +Task | CEFR | NC |
|---|---|---|---|---|---|
| 5 | | | | (C1) | (C1.1) |
| | | | | B2 | B2.1 |
| 4 | 211 | | | | |
| | 5 751 | | | | ---- |
| | 704 830 | | | | |
| 3 | 757 758 761 819 832 | | | --- | |
| | 707 749 750 753 824 | | | | B1.2 |
| | 714 755 756 849 870 873 | | | | |
| 2 | 15 205 871 874 | | | | ---- |
| | 11 17 242 378 387 501 702 717 723 752 811 816 868 9 | | | B1 | B1.1 |
| | 13 21 251 268 320 705 730 739 815 825 872 879 | | | | |
| 1 | 204 241 350 383 386 740 766 821 834 836 839 | | | | ---- |
| | 1 207 217 272 382 390 500 505 741 810 817 820 869 | R3 R6 | | --- | |
| | 260 269 3 721 835 847 | R7 | T1_NC T3_NC | | A2.2 |
| * 0 * | 203 206 246 363 502 7 738 813 818 838 841 855 865 | * R2 R4 * | T1_CEFR T3_CEFR T5_CEFR T2_NC T4_NC T5_NC * | * | * |
| | 234 243 261 266 300 324 325 344 354 372 379 380 504 516 716 726 828 831 843.. | R9 | T2_CEFR T4_CEFR | | ---- |
| | 232 240 245 327 348 364 389 529 733 747 850 861 | R1 | | | |
| -1 | 209 212 214 216 219 221 271 333 384 385 392 833 840 | R8 | | A2 | |
| | 201 213 215 222 235 244 250 303 308 331 355 358 365 521 764 842 862 | | | | A2.1 |
| | 202 230 231 263 335 377 535 856 | | | | |
| -2 | 210 236 264 336 388 391 514 525 718 822 | | | | ---- |
| | 218 252 321 511 524 881 | | | --- | |
| | 200 267 302 332 357 360 369 393 | | | | |
| -3 | 220 237 238 247 248 312 337 339 515 533 880 | | | | A1.3 |
| | 249 270 305 322 362 867 | | | | |
| | 239 274 301 503 506 507 513 522 530 534 | | | | ---- |
| -4 | 307 311 326 342 345 356 508 523 | | | | |
| | 309 316 346 | | | | |
| | 273 313 323 329 334 509 531 | | | A1 | A1.2 |
| -5 | 306 314 510 532 | | | | ---- |
| | 262 265 304 366 527 | | | | |
| | 317 341 | | | | |
| -6 | 526 | | | | |
| | 368 | | | | A1.1 |
| | | | | --- | |
| -7 | | | | | ---- |
| | | | | | |
| | 512 | | | | |
| -8 | | | | (0) | (0) |
| Measr | -Examinee | -Rater | -Task | CEFR | NC |

APPENDIX 4

Finnish CEFLING ratings - CEFR and NCC scales Table 6.0 from Facets output

| Measr | +Examinee | +Rater | +Task | CEFR | NC |
|---|---|---|---|---|---|
| 5 | | | | (C1) | (C1.1) |
| | | | | | B2.2 |
| | 797 | | | | |
| | | | | | |
| | | | | | |
| 4 | | | | B2 | ---- |
| | | | | | |
| | | | | | B2.1 |
| | | | | | |
| | 796 812 | | | | |
| 3 | 29 | | | --- | ---- |
| | 15 246 730 788 794 806 816 | | | | |
| | 20 746 | | | | B1.2 |
| | 35 57 6 711 739 | | | | |
| | 232 28 708 800 809 | | | B1 | ---- |
| 2 | 14 234 27 303 41 704 743 745 817 | | | | |
| | 21 240 50 7 748 787 811 813 824 | | | | |
| | 248 32 707 712 760 766 767 795 808 821 823 | | | | B1.1 |
| | 48 700 706 741 747 783 799 801 | R11 | | | |
| | 1 208 25 49 54 62 731 732 736 765 780 819 | | | | |
| 1 | 11 19 23 710 756 761 814 | | | --- | ---- |
| | 206 233 236 307 31 55 724 733 737 774 789 790 804 818 9 | | | | |
| | 2 205 231 235 243 304 306 308 51 61 713 791 805 807 | | | | |
| | 34 721 740 742 744 755 768 773 782 785 802 803 815 820 | | T1_CEFR T2_CEFR | | A2.2 |
| | 10 16 17 22 241 247 33 47 727 729 734 750 764 784 792 793 798 822 | R6 R3 R7 | T3_CEFR | | |
| * 0 * | 202 24 242 302 717 719 722 738 749 763 | * R9 | * T1_NC T2_NC T3_NC | * | * * |
| | 13 203 237 239 250 3 30 751 778 779 810 | R4 R2 R10 | T5_CEFR T4_NC | | ---- |
| | 200 238 244 26 300 44 63 702 728 735 757 762 772 | R1 | T4_CEFR T5_NC | A2 | |
| | 4 705 752 758 | R8 R5 | | | |
| | 36 42 5 709 714 769 775 786 | | | | |
| -1 | 12 204 45 723 754 771 776 781 | | | | A2.1 |
| | 720 | | | | |
| | | | | | |
| | 301 38 46 59 715 716 725 759 770 | | | | |
| | 207 43 60 753 777 8 | | | | ---- |
| -2 | 245 39 52 726 | | | --- | |
| | 305 58 701 | | | | |
| | 40 | | | | |
| | 201 | | | | A1.3 |
| | 18 | | | | |
| -3 | 249 | | | A1 | |
| | 53 | | | | |
| | 703 718 | | | | ---- |
| | 37 | | | | |
| | 56 | | | | |
| -4 | | | | (0) | (0) |
| Measr | -Examinee | -Rater | -Task | CEFR | NC |

APPENDIX 5

The structure of the CEFR and NCC scales

(Note: 'Used' = number of scripts / writing performances)

CEFR scale

| | English | | Rasch-Andrich Thresholds | | Finnish | | Rasch-Andrich Thresholds | |
|---|---|---|---|---|---|---|---|---|
| Score | Used | % | Measure | S.E. | Used | % | Measure | S.E. |
| Below A1 | 88 | 3 % | | | 3 | 0 % | | |
| A1 | 999 | 29 % | -7,84 | 0,13 | 386 | 14 % | -9,12 | 0,61 |
| A2 | 1369 | 39 % | -2,90 | 0,06 | 1352 | 51 % | -2,20 | 0,07 |
| B1 | 802 | 23 % | 0,70 | 0,06 | 734 | 28 % | 1,70 | 0,05 |
| B2 | 204 | 6 % | 3,62 | 0,09 | 166 | 6 % | 4,09 | 0,09 |
| C1 | 13 | 0 % | 6,41 | 0,29 | 27 | 1 % | 5,52 | 0,21 |

NCC scale

| | English | | Rasch-Andrich Thresholds | | Finnish | | Rasch-Andrich Thresholds | |
|---|---|---|---|---|---|---|---|---|
| Score | Used | % | Measure | S.E. | Used | % | Measure | S.E. |
| Below A1.1 | 27 | 1 % | | | 2 | 0 % | | |
| A1.1 | 89 | 3 % | -5,65 | 0,22 | 13 | 0 % | -5,54 | 0,72 |
| A1.2 | 360 | 11 % | -5,01 | 0,11 | 89 | 3 % | -5,03 | 0,28 |
| A1.3 | 494 | 15 % | -2,99 | 0,07 | 247 | 9 % | -3,26 | 0,12 |
| A2.1 | 728 | 23 % | -1,96 | 0,06 | 809 | 31 % | -2,41 | 0,07 |
| A2.2 | 657 | 21 % | -0,41 | 0,05 | 641 | 24 % | -0,07 | 0,05 |
| B1.1 | 423 | 13 % | 0,86 | 0,06 | 509 | 19 % | 0,73 | 0,06 |
| B1.2 | 274 | 9 % | 1,69 | 0,07 | 211 | 8 % | 2,15 | 0,08 |
| B2.1 | 132 | 4 % | 2,73 | 0,10 | 73 | 3 % | 3,08 | 0,13 |
| B2.2 | 16 | 0 % | 4,77 | 0,26 | 20 | 1 % | 4,24 | 0,26 |
| C1 | 1 | 0 % | 5,96 | 1,01 | 3 | 0 % | 6,11 | 0,65 |

APPENDIX 6

Task performance across the scales and languages

Task performance on the CEFR scale

(for the Observed and Fair-M averages, A1=1, A2=2, B1=3, B2=4)

| | English | | | | | | | | | Finnish | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | Total Score | Total Count | Obs. Aver. | Fair-M Avrage | Measure (Model S.E.) | Infit MnSq | ZStd | Outfit MnSq | ZStd | Total Score | Total Count | Obs. Aver. | Fair-M Avrage | Measure (Model S.E.) | Infit MnSq | ZStd | Outfit MnSq | ZStd |
| **Task1** | 986 | 470 | 2,1 | 1,91 | -0,05(0,09) | 0,92 | -1,20 | 0,91 | -1,20 | 731 | 317 | 2,3 | 2,26 | 0,30(0,10) | 0,81 | -2,30 | 0,77 | -2,60 |
| **Task2** | 980 | 492 | 2,0 | 1,90 | -0,09(0,09) | 1,07 | 0,90 | 1,05 | 0,70 | 877 | 356 | 2,5 | 2,28 | 0,35(0,09) | 0,99 | 0,00 | 1,00 | 0,00 |
| **Task3** | 1729 | 844 | 2,0 | 2,01 | 0,37(0,07) | 0,98 | -0,30 | 0,97 | -0,50 | 1518 | 670 | 2,3 | 2,24 | 0,20(0,07) | 1,13 | 2,10 | 1,11 | 1,70 |
| **Task4** | 1679 | 845 | 2,0 | 1,87 | -0,21(0,07) | 1,00 | 0,00 | 1,02 | 0,20 | 1434 | 665 | 2,2 | 2,07 | -0,45(0,07) | 0,99 | -0,20 | 0,97 | -0,50 |
| **Task5** | 1656 | 828 | 2,0 | 1,91 | -0,02(0,07) | 0,98 | -0,40 | 0,99 | -0,10 | 1531 | 660 | 2,3 | 2,09 | -0,39(0,07) | 0,94 | -1,00 | 0,95 | -0,70 |
| Mean (Count: 5) | 1406,0 | 695,8 | 2,0 | 1,92 | 0 (0,07) | 0,99 | -0,20 | 0,99 | -0,20 | 1218,2 | 533,6 | 2,3 | 2,19 | 0,00(0,08) | 0,97 | -0,30 | 0,96 | -0,40 |
| S. D. (Population) | 346,2 | 175,6 | 0,0 | 0,05 | 0,20(0,01) | 0,05 | 0,70 | 0,05 | 0,70 | 343,0 | 161,4 | 0,1 | 0,09 | 0,35(0,01) | 0,10 | 1,50 | 0,11 | 1,40 |
| S. D.(Sample) | 387,1 | 196,4 | 0,0 | 0,06 | 0,22(0,01) | 0,05 | 0,80 | 0,05 | 0,80 | 383,4 | 180,5 | 0,1 | 0,10 | 0,39(0,01) | 0,11 | 1,60 | 0,12 | 1,60 |

Task performance on the NCC scale

(for the Observed and Fair-M averages, A1.1=1, A1.2=2, A1.3=3, A2.1=4, A2.2=5, B1.1=6, B1.2=7, B2.1=8, B2.2=9, C1.1=10)

| | English | | | | | | | | | Finnish | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | Total Score | Total Count | Obs. Aver. | Fair-M Avrage | Measure (Model S.E.) | Infit MnSq | ZStd | Outfit MnSq | ZStd | Total Score | Total Count | Obs. Aver. | Fair-M Avrage | Measure (Model S.E.) | Infit MnSq | ZStd | Outfit MnSq | ZStd |
| **Task1** | 1960 | 418 | 4,7 | 4,46 | 0,18(0,05) | 0,89 | -1,60 | 0,90 | -1,50 | 1499 | 309 | 4,9 | 4,99 | 0,13(0,07) | 1,13 | 1,50 | 1,13 | 1,40 |
| **Task2** | 1554 | 367 | 4,2 | 4,20 | -0,17(0,06) | 1,11 | 1,40 | 1,11 | 1,40 | 1774 | 349 | 5,1 | 4,97 | 0,10(0,06) | 1,03 | 0,40 | 1,00 | 0,00 |
| **Task3** | 3862 | 851 | 4,5 | 4,53 | 0,27(0,04) | 1,07 | 1,40 | 1,05 | 1,00 | 3104 | 649 | 4,8 | 4,93 | 0,05(0,05) | 1,12 | 2,00 | 1,11 | 1,90 |
| **Task4** | 3692 | 845 | 4,4 | 4,18 | -0,19(0,04) | 1,08 | 1,50 | 1,08 | 1,60 | 3138 | 662 | 4,7 | 4,83 | -0,09(0,05) | 0,93 | -1,20 | 0,93 | -1,20 |
| **Task5** | 3097 | 724 | 4,3 | 4,27 | -0,08(0,04) | 0,81 | -3,80 | 0,82 | -3,50 | 3183 | 648 | 4,9 | 4,75 | -0,19(0,05) | 0,90 | -1,70 | 0,91 | -1,60 |
| Mean (Count: 5) | 2833,0 | 641,0 | 4,4 | 4,33 | 0,00(0,05) | 0,99 | -0,20 | 0,99 | -0,20 | 2539,6 | 523,4 | 4,9 | 4,90 | 0,00(0,05) | 1,02 | 0,20 | 1,02 | 0,10 |
| S. D. (Population) | 923,5 | 208,5 | 0,2 | 0,14 | 0,19(0,01) | 0,12 | 2,20 | 0,11 | 2,00 | 742,9 | 159,3 | 0,1 | ,09 | 0,12(0,01) | 0,09 | 1,50 | 0,09 | 1,40 |
| S. D.(Sample) | 1032,5 | 233,1 | 0,2 | 0,16 | 0,21(0,01) | 0,13 | 2,40 | 0,13 | 2,30 | 830,6 | 178,1 | 0,1 | ,10 | 0,14(0,01) | 0,10 | 1,70 | 0,10 | 1,60 |