Comparing the outcomes of two different approaches to CEFR-based rating of students' writing performances across two European countries

Abstract

This study investigated to what extent two teams of experienced raters from different European countries (Finland and Austria), using their own CEFR-based rating scale (one holistic and one analytic), agreed on the CEFR level of students' writing performances. Both teams rated one hundred performances written by Austrian secondary school students based on two tasks. The Finnish raters (N=3) applied a holistic CEFR-linked rating scale consisting of verbatim CEFR descriptors developed in Finland, while the Austrian team (N=6) used an analytic CEFR-linked rating scale consisting of four criteria developed in Austria. The ratings were analysed using the Rasch model.

Although there were individual differences in rater severity among both teams of raters, a clear pattern emerged from the data: The Austrian raters were slightly more lenient than the Finnish raters. Although there was a statistically significant difference in rater severity between the two groups, the actual scope of disagreement was small. Thus, overall, the two teams agreed to a large extent on the CEFR levels of the participants.

Keywords

CEFR levels; assessing writing; rating scales; rater effects; Rasch model

1. Background

Since its publication in 2001 the Common European Framework of Reference (CEFR) for Languages has had a lasting impact on language education in general and on language assessment in particular (Council of Europe, 2006; Deygers, Zeidler, Vilcu, & Carlsen, 2017; Harsch, 2016; Jones & Saville, 2009; Little, 2007; Martyniuk & Noijons, 2007). Its widespread use as underlying construct for language assessment systems in an increasing number of countries has begged the question whether CEFR-based results are comparable across different contexts. The same test outcomes for exams based on the same ~~of the~~ CEFR descriptors "is clearly a goal worth pursuing, for purposes within education and beyond it [, because] particular contexts or particular languages may refer the [CEFR] level descriptors to different realities, and thus interpret them differently" (European Commission, 2012, p. 21). However, surprisingly, studies addressing this issue are sparse, also with regards to the use of CEFR descriptors for writing assessment.

Within specific national contexts, the use of CEFR descriptors for writing assessment and rating scale development has been investigated by a number of studies. Many of these investigations seem to relate to second language acquisition (SLA) research and to the use of learner corpora in SLA studies. One of the issues in SLA studies is how to determine the proficiency level of the learners and their performances that are being investigated because knowing learners' proficiency with some precision can help interpret the findings of the studies. The wide use of the CEFR has made its proficiency scale a very appealing tool as it offers a more precise and reliable way of finding out the stage at which the learners of interest are than using generic categorisations of learners such as 'beginners', 'intermediate' or 'advanced'. Thus, the focus of studies that have used CEFR scales for rating second (L2) or foreign (FL) language learners' performances has not been to investigate the comparability of the CEFR scale in different contexts, but to use it as a practical tool to improve the quality of the placements of learners or their performances on a scale that is more transparent than other systems used to describe learners' proficiency. However, such research has the potential, as a side product, to provide us with information about the success of applying CEFR scales for rating performances. For example, if the researchers manage to achieve fairly high levels of inter-rater reliability when applying the CEFR scale, this suggests that the scale descriptors are informative enough to allow experienced raters familiar with the CEFR to use them consistently in the same way. More to the point of the current study, if such studies involve two or more contexts (languages, countries, etc.) and if the researchers report on the similarities or differences between them, we may draw evidence about the extent to which the CEFR scales lead to common classifications of learners in different contexts.

Two studies within an SLA context were conducted by Carlsen (2010, 2012), who reports on the creation of a corpus of Norwegian as L2 writing performances based on 1,222 texts, each rated by 5 to 10 raters. The ratings were based on nine CEFR scales, including overall written production, reports and essays, creative writing, and several more specific or linguistically oriented CEFR scales such as general linguistic range, vocabulary range and control, and coherence and cohesion (Carlsen 2012, p. 173). Rater reliabilities were calculated in several different ways and they all turned out to be relatively high. This well designed large-scale study is interesting as it showed that it is possible to achieve a shared understanding of the meaning of the CEFR levels in one context if careful attention is given to rater training. However, the study does not shed light on how comparable the Norwegian judges' ratings are with other pools of raters in other countries.

In a different SLA study, Kuiken, Vedder and Gilabert (2010) investigated the relationship between communicative adequacy and linguistic complexity in second and foreign language writing. A total of 34 international L2 learners of Dutch, 42 Dutch FL learners of Italian and 27 Dutch FL learners of Spanish completed two writing tasks. The researchers used general descriptors from the CEFR, in an adapted form, for rating learners' performances on a 6-point scale. The authors did not explicate if the scale was intended to match the 6-point CEFR levels. Inter-rater reliabilities between .700 and .882 (Cronbach's alpha) were achieved, depending on the language and dimension that was rated; of the two dimensions, linguistic complexity was somewhat more reliably rated than communicative adequacy. Although the study covered three languages, and both L2 and FL learners, the authors did not report in detail on the quality of the ratings across the contexts since other questions were investigated.

Within the field of language testing, a small number of studies also focussed on the use of the CEFR for rating purposes. Harsch and Martin (2012) validated a CEFR-based rating scale for written performances in the German secondary school context. They report that their combined rater training and scale revision approach enabled them to adapt the CEFR descriptors to fit the local context in a way that led to more reliable and valid ratings. Harsch and Martin also undertook a sorting exercise of the finalized scale descriptors into levels and criteria with a group of 14 external experts and report relatively high levels of agreement. However, they did not investigate how their CEFR-based ratings compared to those of other contexts.

Authors (2014) looked at Finnish researchers' ratings of approximately one thousand L2 writing performances by Finnish students. The ratings in this study, that combined language testing and SLA perspectives, were based on a holistic rating scale consisting of verbatim CEFR descriptors and the authors report that the scale was applied in a consistent way and with all the levels of the scale being separable from the adjacent levels. When comparing the holistic CEFR scale ratings with ratings on a more fine-grained CEFR-based scale, the authors found that the ratings corresponded closely in terms of CEFR levels. Only at the lower levels the holistic verbatim CEFR ratings tended to be more lenient. Authors (2014) argue that although the Finnish raters applied the scales reliably, "we do not know how representative our view is of the meaning of the [CEFR] levels […], as it has not been possible to compare our assessments with those by other groups of raters in other countries" (Authors, 2014).

Finally, in a study by Deygers and Gorp (2015) a CEFR-based rating scale, which was co-constructed by Dutch-speaking novice raters, was validated in terms of rater reliability and uniformity of descriptor interpretation. Six Dutch-speaking novice raters first applied the scale to 200 performances (100 speaking performances and 100 writing performances) and then took part in a focus group discussion on their interpretation of the scale descriptors. After analysing the results Deygers and Van Gorp conclude that although the ratings were statistically reliable, "achieving a uniform interpretation of CEFR-based descriptors remains a challenge" (2015, p. 537). The raters in this study perceived the CEFR levels as too broad and multifaceted for some criteria and certain descriptors as too vague, however, all raters were again from the same country and working in the same context.

In sum, although the limited number of studies available on this topic outlined above have addressed how CEFR descriptors can be used for or adapted for rating purposes of written performances in order to achieve reliable ratings, research into how comparable CEFR-based ratings of written

performances are across national and educational contexts is still lacking. The current study attempts to start to fill this gap by addressing the following research question:

To what extent do two teams of raters from different European countries, using their own CEFR-based rating scale (one holistic and one analytic), agree on the CEFR levels of students' writing performances?

2. Data and Methods

2.1. Tasks

Two tasks were used in the study. Both tasks were developed and used for research purposes in the Finnish context and have been shown to work well (Authors, 2012; Authors, 2014). One task was an email of complaint and the other task an opinion essay (see Appendix A). The tasks were translated into German for the Austrian students in order to avoid potential misunderstandings and to minimize lifting of input from the prompt.

2.2. Performances

One hundred student performances were collected for the study. The performances were based on the two tasks (50 for each task) and were written by Austrian secondary school students. The Finnish tasks were used in Austria because a) the tasks have been shown to work well for research purposes and b) a suitable number of Austrian students was available at the time of data collection. Performances were collected during students' classes by their class teachers under conditions that were similar to those under which tests are typically administered. Class teachers received detailed instructions for test administration. A reasonable amount of time was given to the students to complete the tasks given to them.

The students were between 15 and 19 years old and were equally spread in terms of grades, schools, and regions across Austria (overall, performances were collected across all four grades in eight different academic upper secondary schools in seven different regions of Austria). The students' English language proficiency in terms of CEFR level, as judged by the Austrian upper secondary curriculum, was expected to be in the B1 and B2 range, as students in the first two grades of upper secondary school should be at CEFR B1, and in the two final years of their education should be at CEFR B2.

2.3. Rating contexts

Austria and Finland are countries with a somewhat different history in applying the CEFR in their educational systems, which may lead to differences in test outcomes.-At the very least, we cannot take similar-test outcomes for granted. Finland has been involved in the development of CEFR-related materials for many years and has developed training tools such as CEFTrain (http://www.ceftrain.net). Interestingly, the impact of the CEFR on language learners' assessment in Finland is mostly felt in adult education and in research studies (see Authors, 2008). While the National Curriculum for primary and secondary education in Finland has included CEFR-related scales since the early 2000s (Finnish National Board of Education, 2003), the scales mainly illustrate learning and teaching goals and have limited use in assessment: teachers' marking is only indirectly influenced by them and the final school leaving examination (the Matriculation Examination) does not use the CEFR at all. In contrast, in Austria the CEFR has been largely introduced into the school system

through a government-funded national exam reform (Authors, 2016) using familiarization materials which were specifically developed for the reform and which were tailored to the specific needs of the stakeholders involved. The CEFR was at the heart of this reform, in the course of which standardized language exams have been developed across four languages and two CEFR levels.

2.4. Rating scales

Two CEFR-based rating scales for grading students' writing performances were used in the study, one holistic and one analytic. The holistic rating scale was developed for grading Finnish EFL learners' texts for research purposes. The analytic rating scale was developed for the standardized Austrian school leaving examination (Matura).

The Finnish rating scale has been used for several research projects in Finland since 2007 (Authors, 2012; Authors, 2014). It is based on seven writing-related CEFR scales, which were taken verbatim from the CEFR and are grouped into four general areas: (1) overall written production (Council of Europe, 2001, p. 61); (2) written interaction (p. 83); (3) correspondence (p. 83); and notes, messages, forms (p. 84); and (4) creative writing (p. 62); thematic development; and coherence and cohesion (p. 125). The scale is divided into six bands, one for each CEFR level. The raters award one overall CEFR rating to each performance (see Authors, 2014 for more details of the scale).

The CEFR-linked Austrian Assessment Scale at B2 (Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens, 2014) is used for rating Austrian students' English writing performances for classroom assessment in the penultimate and ultimate year of secondary education and for rating writing performances in the school leaving exam (Matura). The scale was developed over a three year period by a team of language testing professionals, item writers, and international experts (for a detailled outline of the scale development process see Authors, in press). It is an analytic rating scale, consisting of four criteria: task achievement, organization and layout, lexical and structural accuracy, and lexical and structural range.

All four criteria are divided into 11 bands (0 to 10), with 6 bands containing descriptors and 5 bands without descriptors, for performances which lie within two described bands. The scale developers stipulated band 6 as the pass mark. As the English exam is aimed at CEFR B2, bands 6, 8, and 10 contain B2 descriptors which have either been taken verbatim from the CEFR or have been modified to allow for sufficiently detailed gradation towards the higher levels. Correspondingly, bands 0, 2, and 4 contain B1 descriptors and modifications thereof. The described bands also contain additional descriptors not taken from the CEFR, which were written during the scale development process. For example, the criterion task achievement contains a number of descriptors specifically tailored to the tasks used in the Austrian Matura exam, such as task type requirements, number of content points, or set word length. For this reason, a small number of descriptors in the criteria task achievement and organization and layout had to be deleted or adapted slightly for use with the tasks in this study. Each of the four criteria is rated separately, so students could achieve band 8 in task achievement but at the same time band 4 in lexical and structural accuracy.

In the Matura exam students do not receive individual CEFR scores for each of the four areas assessed (reading, listening, writing, and lexico-grammar) but only an overall CEFR score. That is because students can compensate their lack of proficiency in one skill with their proficiency in another, although there is a minimum requirement for each skill. However, according to the guidelines on grading classroom assessments with the Austrian rating scale published by the Austrian

Ministry of Education, students need to achieve at least band 6, or 24 "points" overall across the four criteria (as they again have the chance to compensate for their lack of proficiency in one criterion with their proficiency in another), to be awarded B2 for writing alone (Bundesministerium für Bildung, 2015; Bundesministerium für Bildung und Frauen, 2013, p. 21). Therefore, for the purpose of this study, a pass mark of 24 was used to differentiate between B1 and B2 performances as judged by the raters.

2.5. Raters

Two teams of raters with extensive experience both in using CEFR-based rating scales and in rating English learners in secondary level education took part in the study, one from Finland and one from Austria. The Finnish raters applied the holistic Finnish rating scale to the performances, while the Austrian team used the analytic Austrian scale.

The Finnish team consisted of three raters, all of whom were non-native speakers of English working at a university. All of them were specialists in language education and applied linguistics with at least Masters' level degrees, and with varying amounts of experience in language teaching and assessment, including assessment of EFL learners of the same age as those involved in this study by using CEFR-based scales. One of the raters had participated in all three research projects in which the CEFR scale and the two tasks used in the current study had been used (and also had extensive rating experience in other contexts). The second rater had been an assessor in one of the previous projects and had rated several hundred performances in the course of this project. The third was less experienced, but had rated about 200 performances in another context with the CEFR scale used in the present study. The second and third raters were chosen for two reasons: their work load allowed them to take on the extra task of rating the Austrian performances and the analyses carried out on their earlier ratings indicated that they could rate consistently and that their severity was around the average severity of the other Finnish raters who had worked for the previous projects. All three raters completed a pre-rating training exercise with performances collected in the previous projects on the two tasks (see Authors, 2014 for more information). It should be mentioned that only the first and most experienced rater had earlier participated in rater training with international writing samples illustrating the CEFR levels (i.e., with CEFTrain and writing samples published by the Council of Europe). The (re)training of raters for the current study made use of benchmark samples of performances on the specific tasks used in the study.

The Austrian rater team was made up of specialised secondary school teachers and teacher trainers who had been trained as double-raters for the Austrian Matura exam. The rating of student performances for the Matura still lies in the hands of the individual class teachers, most of whom have not received any formal training in standardized rating of writing performances (Authors, 2016). Therefore, a pool of 32 double-raters was established, which teachers or headmasters can call upon when this is deemed necessary or when there is an appeal. These double-raters have undergone extensive training in the CEFR as well as in applying the Austrian rating scale. They took part in three five-day training courses over the course of one year and have rated hundreds of student performances in their functions as double-raters, language teachers and teacher trainers. Their performance as raters was monitored throughout the course of the training (for details on the training program see Authors, in press). Out of this pool, six teachers were chosen to take part in the study. These six double-raters had been performing best in terms of inter- and intra-rater reliability statistics during and after the training sessions, i.e. their mean ratings and infit mean squares were

consistently good throughout the training (for details see Authors, in press). The six raters were experienced secondary school teachers of English who had studied language education and applied linguistics with at least Masters' level degrees. At the time of data collection for this study, all six raters had been working at teacher training colleges across Austria to familiarize classroom teachers with the CEFR and the Austrian scale. In addition, several of the raters are regularly invited as standard setting judges for the Austrian Matura exam due to their extensive knowledge of the CEFR.

In sum, although the work context of the two teams of raters was different (researchers at a University vs. secondary school teachers and teacher trainers) the level of CEFR familiarity was very high for both groups. With the exception of one of the Finnish raters, both teams had been trained in and working with the CEFR descriptors, and with secondary school level EFL learners, for many years in different national assessment contexts.

2.6. Rating design

Due to the different numbers of raters in the two countries and the different amount of time it takes to rate holistically as compared to analytically, different rating designs had to be used for the two teams of raters. While the Finnish raters assessed all 100 performances collected in Austria, for the Austrian raters a different design was implemented. Out of the 100 performances, 80 (40 for each task) were each rated by three raters in an overlapping design. The remaining 20 performances (10 for each task) were rated by all six raters. Overall, each rater assessed 60 performances across all four criteria on the 11-band scale. In addition to the scale ratings, the Austrian team was also asked to indicate holistically which overall CEFR level they would award to each performance, based on their understanding of the CEFR rather than their scale ratings. Thus, overall, we collected three different sets of data: The Finnish holistic CEFR ratings, the Austrian analytic ratings, and the Austrian holistic CEFR ratings.

3. Analysis

To identify the extent to which the two groups of raters agreed on the CEFR level of students' writing performances, we coded all ratings as "below B2" or "at or above B2". A total score of 24 or more on the analytic Austrian scale was coded as "at or above B2" (following the guidelines provided in Bundesministerium für Bildung, 2015; Bundesministerium für Bildung und Frauen, 2013, p. 21); the cut-off score for the holistic CEFR Finnish and Austrian ratings was 4 (i.e. B2) or more. Ratings below B2 were coded as 0 and ratings at or above B2 were coded as 1. The analysis was run twice: we compared the Finnish holistic CEFR ratings both with the Austrian scale ratings as well as the Austrian holistic CEFR ratings.

For the study of rater effects researchers often use the Many-Facets Rasch model (Linacre, 1994) which is conceptually a generalization of the simple Rasch model (Rasch, 1960) and allows the direct analysis of many facets, such as examinees, items and raters. However, for the purpose of this study, we decided to analyse each of the two tasks independently, which allowed us to use the simple Rasch model to investigate both the severity of the raters and the distribution of the residuals for each of the two tasks.

The simple Rasch model, with dichotomous scoring (0 or 1), can be described by the following equation:

$$\log\left(\frac{P_{nj}}{1-P_{nj}}\right) = B_n - C_j$$

Where P_{nj} is the probability of student n being assigned a rating of B2 or above by the j^{th} rater on a specific task, C_j is the severity of rater j, and B_n is the ability estimate of student n.

We evaluated the model-data fit using the Infit Mean Square and the Outfit Mean Square (Wright and Stone, 1979). In many studies the range of acceptance for examinee and rater Infit Mean Squares is 0.6 to 1.5, however, the standardized Infit and Outfit Mean Squares need to be within -2 and 2 (see Engelhard, 1992, 1994; Lunz, Wright, & Linacre, 1990). For more details on the evaluation of model-data fit for the Rasch models, see Author (2004) or Author (2006).

For the analysis of the data, we cross-checked the results from the eRm package (Mair, 2016) and the TAM package (Robitzsch, 2017) on the R platform (R Core Team, 2017). To compare the mean severity of the sub-groups of Finnish and Austrian raters, we used the "population model" of the TAM package and confirmed the findings using the lme4 package (Bates et al., 2015) according to the examples suggested by Author (2013). Using this command, the model assumes that the Austrian and the Finnish raters come from different sub-populations of raters with normally distributed severities and that the mean and the standard deviation of the severities of the raters may differ between the two countries.

4. Results

Each of the three Finnish raters rated 100 scripts whereas each of the six Austrian raters rated 60 scripts (overall, there were only five missing ratings). The distribution of raw data (combined for both tasks) is shown in Appendix B.

Initially, as described above, we coded all the ratings dichotomously, as 0 (i.e., "below B2") or 1 (i.e., "at or above B2"). A total score of 24 or more on the analytic Austrian scale was coded as "at or above B2"; the corresponding cut-off score for the holistic CEFR ratings was 4 or more. The analysis was repeated separately for the two different datasets (Task 1 and Task 2).

4.1. Finnish holistic CEFR ratings and Austrian scale ratings

For Task 1, the Rasch analysis shows that some Finnish raters were substantially more severe compared to some of the Austrian raters (see Figure 1, left) and as a result they were less likely to award a rating of B2 or above. More specifically, the Finnish raters 7 and 9 were substantially more severe than the Austrian raters 3, 4 and 5. The model-data fit of the analysis was appropriate for the purposes of the study, according to the guidelines described above. Infit Mean Squares ranged from 0.7 to 1.3 (see also Appendix F). The EAP Reliability index was 0.80, which suggests that the ability of the students was measured with adequate precision.

For Task 2, the Rasch analysis shows that all the three Finnish raters were more severe compared to the Austrian raters (see Figure 1, right) and as a result the Finnish raters were less likely to award a rating of B2 or above. Similar to Task 1, the model-data fit of the analysis was satisfactory, with none of the raters showing misfit (see Appendix G). The EAP Reliability index was 0.80, which again indicates adequate precision in the student ability measures.

In order to investigate whether different cut-off scores for the Austrian analytic sale could have an impact on students' results, we ran a Rasch analysis for a number of different cut-off scores (for detailed examples with relevant discussion, see Author, 2006, 2008). We coded the data as "0" if a rater awarded B1 or less and "1" if a rater awarded B2 or more. For each analysis, we estimated the proportion of students who were awarded a level of B2 or more (the proportion only changed for the Austrian raters, as the cut-off score only changed for their analytic scale). Figure 2 shows that with an increase in the cut-off score on the Austrian analytic scale for Task 1, the proportion of students classified at B2 or above decreases. At a cut-off of 26, the proportion of students classified at B2 or above is almost the same between the two groups of raters. For Task 2, the results are similar; the difference between the two countries regarding the proportion of students classified at B2 or above decreases rapidly while the cut-off score increases. At a cut-off of 27, the two groups of raters assign almost the same number of students to B1 and B2 (Figure 3) and the difference is no longer statistically significant ($\chi^2(1)=1.74$, $p=0.187$).

We also calculated the mean severity of the Austrian raters compared to the Finnish raters for the different cut-off scores. For Task 1, we found that at a cut-off score of 26 for the Austrian analytic scale, the perceived mean severity of the Austrian raters (compared to the Finnish raters) reaches zero. In other words, at a cut-off score of 26, the mean severity of the Austrian and Finnish raters becomes statistically indistinguishable, as shown in Figure 4. For Task 2, the mean severity of the Austrian raters (compared to the Finnish raters) reaches zero at a cut-off of 27 (Figure 5).

4.2. Finnish holistic CEFR ratings and Austrian holistic CEFR ratings

We repeated the analysis comparing the holistic CEFR ratings of both teams of raters to investigate whether the slightly more lenient ratings by the Austrian team might be an artefact of using the analytic rating scale.

The results of the Rasch analysis (Figure 6) echoed the results outlined above. Significant differences were found between the mean severities of the Austrian and Finnish raters (mean severity for the Austrians=-0.79 and mean severity for the Finnish=0.32). These results are very similar to the Task 1 Rasch analysis presented in the previous section (mean severity for the Austrians=-0.82 and mean severity for the Finnish=0.34). The correlation between the rater severity estimates (for Task 1) for the analytic and holistic Austrian scale is $r(7)=0.87$, $p=0.002$.

The results for Task 2 are also very similar. Significant differences were found between the mean severities of the Austrian and Finnish raters (mean severity for the Austrians=-1.16 and mean severity for the Finnish=0.60). There was perfect agreement between the Rasch severity estimates (Task 2) for the analytic and holistic Austrian scale; for all Austrian raters and for all students, a holistic score of 4 or more corresponded to an analytic score of 24 or more.

For the holistic CEFR ratings, the correlation between the rater severity estimates (for all the raters) for Task 1 and Task 2 is very high; $r(7)=0.71$, $p=0.03$.

5. Discussion

One of the main goals of the CEFR is to "[provide] a common basis for the elaboration of language [...] examinations [...] across Europe" (Council of Europe, 2001, p. 1). Therefore, using the CEFR descriptors for rating student performances should ideally lead to the same results across different national assessment contexts. However, research addressing this is sparse, also with regards to the

CEFR writing descriptors. It seemed prudent to investigate whether the use of CEFR descriptors used for rating students' writing performances yields the same results across different educational settings, for two main reasons:

1. "Scoring written essays is a fundamentally interpretive and judgmental activity, based on prevailing norms of educational practice as well as individuals' past experiences." (Cumming, Kantor, & Powers, 2001, p. 72)
2. Rating scales comprised of descriptors similar to the ones included in the various CEFR tables have been criticized for their use of "impressionistic terminology which is open to subjective interpretations" (Knoch, 2009, p. 277, citing Brindley, 1998; Upshur & Turner, 1995; and Watson Todd, Thienpermpool, & Keyuravong, 2004).

The CEFR scales are not meant to be used directly for rating purposes. They are proficiency scales and serve a very different purpose than rating scales. However, research has shown that raters who have been familiarised profoundly with the CEFR scales and the meaning of the CEFR levels can apply the CEFR descriptors reliably when rating student performances (Authors, 2012; Authors, 2014).

This study investigated the extent to which two teams of highly experienced raters from different European countries (Finland and Austria), applying their own CEFR-linked rating scales (one holistic and one analytic), agreed on the CEFR levels of 100 writing performances written by Austrian academic upper secondary school students at B1 and B2 level. The results show that there were individual differences in rater severity among both teams of raters, a phenomenon which is well-documented in the literature (see for example Bachman, Lynch, & Mason, 1995; Congdon & McQueen, 2000; Deygers & Van Gorp, 2015; Eckes, 2005, 2008; Engelhard, 1994; Engelhard & Myford, 2003; Lumley & McNamara, 1995; Schoonen, 2005; Weigle, 1998). Despite these differences, a clear pattern emerged from the data: The Austrian raters, who used an analytic CEFR-linked rating scale consisting of four criteria, were slightly more lenient than the Finnish raters, who used a holistic CEFR-linked rating scale consisting of verbatim CEFR descriptors. In other words, the Austrian raters awarded B2 more often than the Finnish raters. These results were confirmed by an additional analysis based on a comparison of the Austrian raters' holistic CEFR ratings (rather than their scale ratings) with the Finnish ratings.

Although there was a statistically significant difference in rater severity between the two groups, the actual scope of disagreement (i.e. the practical difference for real-world use of the assessments) was small (for an informative discussion of practical vs. statistical significance see Kirk, 1996). With an increase in the cut-off score between B1 and B2 on the Austrian scale by only 2 points for Task 1 and 3 points for Task 2 (on a 40-point scale measuring B1 and B2 writing), the disagreements become practically indistinguishable. This suggests that the majority of disagreements concerned borderline candidates (i.e. candidates that were on the border between B1 and B2 in their writing proficiency). Thus, overall, the results show that the CEFR writing scales can lead to comparable classifications across different contexts with raters who are highly familiar with the CEFR descriptors and have been exposed to hundreds of performances in a community of practice. This conclusion is also supported by the fact that the results were almost the same for the two different writing tasks used in the investigation.

The results thus differ from recent findings by Deygers, Gorp and Demeester (2018), who report significant disagreement on the same students' CEFR scores between two different CEFR-linked L2 Dutch speaking exams. At first glance one could suspect that achieving common CEFR scores across

exams or contexts may be more challenging for speaking skills than for writing. However, a number of other factors may also have had an impact on Deygers et al.'s results. First, there are some inconsistencies in the scales of the two exams Deygers et al. compared, which are acknowledged by the authors. For example, for pronunciation, where the authors report significant disagreement in CEFR scores, one of the two scales includes an original B2 descriptor which was "supplemented with a B1 characteristic" (Deygers et al., 2018, p. 52). In addition, the authors compared two teams of raters who differed considerably in terms of rating experience. One group were "experienced L2 teachers of Dutch who typically attend training at least once a year and score oral tests at different times throughout the year", while the other group were "novice raters with a background in linguistics or communication [who] […] undergo a two-day training and take part in a trial rating session to establish their consistency and reliability" (Deygers et al., 2018, p. 47). Related to that, Deygers et al. do not mention the raters' level of CEFR familiarity, which may also have been different between the two groups. This combination of inconsistencies in scale development and very different rater populations may well have contributed to the lack of equivalence between the two exams. Deygers et al's conclusion that the CEFR is "unusable as a standard" because it "intentionally lacks […] exactness" (Deygers et al., 2018, p. 54) can therefore be challenged. Taking the authors' metaphor of comparing the CEFR to a standardized screw thread further, it is not enough to invent the thread, but one also needs to develop tools that fit it, as well as learn how to use the tools. Our results suggest that it is possible to achieve comparable results on students' CEFR scores across educational contexts, if those scores are based on thoroughly developed rating scales and attained by highly experienced and trained raters.

In order to investigate whether the small differences in severity between the Austrian and Finnish raters might be related to the type of scale used, seeing that previous research has shown that analytic scoring can lead to slightly higher scores than holistic scoring (Hunter, Jones, & Randhawa, 1996), the analysis was repeated using the Austrian team's holistic CEFR ratings which were also collected during the study. This analysis yielded the same results, suggesting that scale format might not be a decisive variable.

Apart from these major findings, our study also has implications for standard setting and cut-score definitions, as well as rater training. Using cross-educational, cross-national, and/or cross-contextual comparisons of the kind outlined in this paper would be one further way to validate cut-scores in exams, especially enhancing insights into the borderline regions. This would be particularly beneficial for exams linked to the CEFR, considering the CEFR's widespread use across the globe. In addition, using the graphs on rater leniency/harshness for rater training purposes could be an effective way to make raters aware of their idiosyncrasies, thereby fostering agreement between raters.

Some limitations of our study concern the necessary modifications to the Austrian rating scale, the student sample, and the B2 cut-off point for the Austrian scale. Although the changes to the Austrian scale were minimal and only affected two of the four criteria, the lenience of the Austrian raters might to some extent be related to these modifications. In addition, the Austrian raters might have been more used to certain flaws that are typical for German L1 users of English and might therefore have been more lenient in their ratings. Also, although the Austrian ministry of education sets a cut-off of 24 points overall for the Austrian scale for a student to qualify as B2, no empirical research has yet been published to validate this. However, the Austrian raters' holistic CEFR ratings in our study correlated very highly with their scale ratings and a cut-off of 24 points, thereby indicating that the cut-off might be valid.

It should also be mentioned that the findings of this study do not necessarily generalise to other groups of raters in either of the countries involved in the study. First, both groups represented trained and, in the great majority of cases, highly experienced raters who had used CEFR-related rating scales before and who had formed an understanding of the meaning of the scale levels not only based on the scale descriptors but, importantly, also based on a great number of concrete performances they had encountered during rater training and actual rating sessions. Therefore, the results do not apply to ordinary language teachers who may want to use CEFR scales for rating their students but without access to adequate rater training. As Deygers et al. (2018) have shown in the context of speaking, less experienced raters may not agree to such an extent on students' CEFR levels. Secondly, while the Austrian raters can be argued to represent an "official" Austrian view of the CEFR levels because of the high stakes nature and wide use of that examination in the country, the same cannot be said about the Finnish raters. Although the CEFR scale is used, in a modified form, in the national curricula for basic and secondary education in Finland, the scale is not used in everyday school based assessments by most teachers, neither is it used in the Finnish equivalent of the Austrian Matura examination. Thus, there is no national consensus of the meaning of the CEFR levels in Finland in the same way as Austria is likely to have; the "Finnish view" of the CEFR is therefore limited to the community of researchers mostly based in one university that has been active in applying the CEFR scales to SLA and language testing research.

6. Conclusion

To our knowledge, this study is the first to directly compare CEFR-based ratings of writing performances between raters of different national and educational contexts. The results give us confidence that the use of writing-related CEFR descriptors for rating purposes may indeed yield comparable results among raters in different European countries, if those raters are trained and highly experienced in using thoroughly developed CEFR-based rating scales. However, it was also shown that borderline candidates between B1 and B2 are likely to be rated more favourably by Austrian raters using their analytic scale than by Finnish raters using their holistic scale. It should also be noted that the direct employment of CEFR descriptors for rating purposes requires extended training and experience and cannot be expected from classroom teachers.

While the results are promising, the current study is only a starting point in researching whether the CEFR descriptors allow for common assessment results across educational contexts. Considering the CEFR's increasing use in a wide variety of language assessments across the globe, studies such as this one seem timely. Future research could thus replicate the study's design in other assessment contexts, also focussing on other language skills than writing. In addition, the current study was not able to reveal the processes raters employ to arrive at their final scores (e.g. Lumley, 2005). Future studies might therefore also want to consider collecting qualitative data on how raters interpret scale descriptors while grading.

References

Alanen, R., Huhta, A., Jarvis, S., Martin, M., Tarnanen, M., 2012. Issues and challenges in combining SLA research and language testing. In: Tsagari, D., Csepes, I. (Eds.),

Collaboration in language testing and assessment. Peter Lang, Frankfurt am Main, pp. 15–30.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. Language Testing, 12, 238–

257.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

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

Bundesinstitut für Bildungsforschung Innovation & Entwicklung des österreichischen Schulwesens. (2014). Assessment scale B2 and guidelines. Vienna, Austria. Retrieved from https://www.srdp.at/downloads/?tx_solr%5Bfilter%5D%5B1%5D=subject%253A%252FLebende%2BFremdsprachen&tx_solr%5Bfilter%5D%5B0%5D=documentType%253A%252FBegleitmaterialien%252FRichtlinien%252C%2BKonzepte%2B%2526%2BModelle

Bundesministerium für Bildung. (2015). Hilfsskala 1 für schriftliche Überprüfungen mit gleicher Gewichtung der Aufgabenbereiche. Vienna, Austria. Retrieved from https://www.bmb.gv.at/schulen/unterricht/ba/reifepruefung_ahs_mslf_lf.html

Bundesministerium für Bildung und Frauen. (2013). Leitfaden zur Erstellung von Schularbeiten in der Sekundarstufe 2 – AHS. Vienna, Austria. Retrieved from https://www.bmb.gv.at/schulen/unterricht/ba/reifepr_ahs_mslf_lf.pdf?5te96f

Carlsen, C. H. (2010). Discourse connectives across CEFR-levels: A corpus based study. In I. Bartning, M. Martin, & I. Vedder (Eds.), Communicative proficiency and linguistic development: Intersections between SLA and language testing research. EUROSLA Monograph Series, 1. 191-209. http://eurosla.org/monographs/EM01/EM01home.html.

Carlsen, C. H. (2012). Proficiency level - a fuzzy variable in computer learner corpora. Applied Linguistics, 33, 161–183.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. Journal of Educational Measurement, 37, 163–178.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe. (2006). Survey on the use of the Common European Framework of Reference for Languages (CEFR). Synthesis of results. Strasbourg, France: Council of Europe. Retrieved from http://www.coe.int/T/DG4/Linguistic/Source/Surveyresults.pdf

Cumming, A., Kantor, R., & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph Series, MS-22). Princeton, NJ.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. Language Testing, 32(4), 521–541.

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. Language Assessment Quarterly, 15(1), 44–58. http://doi.org/10.1080/15434303.2017.1421955

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One framework to unite them all? Use of the CEFR in European university entrance policies. Language Assessment Quarterly. Advance online publication. http://doi.org/10.1080/15434303.2016.1261350

Eberharter, K., Kremmel, B., Holzknecht, F., 2018. Evaluating the effectiveness of a training program

for double-raters (in press). Sigott, G. (Ed.), Language testing in Austria: Taking stock (Sprachtesten in Österreich: Eine Bestandsaufnahme). Peter Lang.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch Analysis. Language Assessment Quarterly, 2(3), 197–221. http://doi.org/10.1207/s15434311laq0203

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. Language Testing (Vol. 25). http://doi.org/10.1177/0265532207086780

Engelhard, G., Jr. (1992). The measurement of writing competence with a many-faceted Rasch model. Applied Measurement in Education, 5, 171-191.

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a Many-Faceted Rasch model. Journal of Educational Measurement, 31, 93–112.

Engelhard, G. J., & Myford, C. M. (2003). Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch model (College Board Research Report No. 2003–1). New York: College Entrance Examination Board.

European Commission. (2012). First European survey on language competencies: Final report. Luxembourg: Publications office of the European Union. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf

Finnish National Board of Education (2003). National Core Curriculum for Upper Secondary Schools 2003. Helsinki.

Goodreau, S. (2007). Advances in exponential random graph (p*) models applied to a large social network. Social Networks, 29, 231–248.

Harsch, C. (2017). Proficiency. ELT Journal, 71(2), 250-253. http://doi.org/10.1093/elt/ccw067

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. Assessing Writing, 17(4), 228–250.

Holzknecht, F., Kremmel, B., Konzett, C., Eberharter, K., Konrad, E., Spöttl, C., 2018. Potentials and challenges of teacher involvement in rating scale design for high-stakes exams (in press). Xerri, D., Vella Briffa, P. (Eds.), Teacher involvement in high stakes language testing. Springer.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., Hirvela, 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.

Hunter, D., Jones, R. M., & Randhawa, B. S. (1996). The Use of holistic versus analytic scoring for large-scale assessment of writing. The Canadian Journal of Program Evaluation, 11(2), 61–85. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ543798

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software, 24(3), 1–29.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. Language Assessment Quarterly, 10(2), 135–159. http://doi.org/10.1080/15434303.2013.769545

Jones, N., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. Annual

Review of Applied Linguistics, 29(2009), 51–63. http://doi.org/10.1017/S0267190509090059

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746–759.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. Language Testing, 26(2), 275–304. http://doi.org/10.1177/0265532208101008

Lamprianou, I., 2004. Application of single-level and multi-level Rasch models using the lme4 package. Journal of Applied Measurement 14 (1), 79–90.

Lamprianou, I., 2006. The stability of marker characteristics across tests of the same subject and across subjects. Journal of Applied Measurement 7 (2), 192–200.

Lamprianou, I., 2008. High stakes tests with self-selected essay questions: Addressing issues of fairness. International Journal of Testing 18 (1), 55–89.

Lamprianou, I., Boyle, B., 2004. Accuracy of measurement in the context of mathematics national curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. Journal of Educational Measurement 41 (3), 239–260.

Linacre J.M. (1994). Many-Facet Rasch Measurement. 2nd Ed. Chicago: MESA Press

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), 645–653.

Lumley, T. (2005). Assessing second language writing: The rater's perspective. Frankfurt am Main: Peter Lang.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. Language Testing, 12, 54–71.

Lunz M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 3(4), 331-345.

Mair, P., Hatzinger, R., & Maier M. J. (2016). eRm: Extended Rasch Modeling. 0.15-7. http://erm.r-forge.r-project.org/

Martyniuk, W., & Noijons, J. (2007). Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States. Strasbourg, France: Council of Europe. Retrieved from http://www.coe.int/T/DG4/Linguistic/Source/Survey_CEFR_2007_EN.doc

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Denmarks Paedagogiske Institut.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. Social Networks, 29, 173–191.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules. R package version 2.8-21 https://CRAN.R-project.org/package=TAM

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. Language Testing, 22, 1–30.

Scott, J. (1988). Social Network Analysis. Sociology, 22(1), 109–127.

Spöttl, C., Kremmel, B., Holzknecht, F., Alderson, J.C., 2016. Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. Papers in Language Testing and Assessment 5 (1), 1–22.

Tarnanen, M., Huhta, A., 2008. Interaction of language policy and assessment in Finland. Current Issues in Language Planning 9 (3), 262–281.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. ELT Journal, 49(1), 3–12.

Watson Todd, R., Thienpermpool, P., & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. Assessing Writing, 9, 85–104.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15, 263–287.

Wright, B. D., & Stone, M. H. (1979). Best Test Design Rasch Measurement. Chicago, IL Mesa Press

Appendix

A. The tasks used in the study (the tasks were translated into German for the participants)

Task 1:

Your parents have ordered a computer game OR a piece of clothing (you can choose which product) for you from a British internet store. When you get the game you notice that it doesn't work properly (OR if you chose a piece of clothing, you notice that there is something wrong about it). You get upset and decide to write an email message to the internet store. In the message, say

- who you are
- what your parents ordered
- why you are unhappy (mention at least two defects/problems)
- how you would like them to take care of the matter
- give your contact information

Remember to begin and end the message appropriately. Write in English in clear characters in the space below.

Task 2:

You have read an argument in a newspaper where somebody says that students should not be allowed to bring mobiles phones to school. You decide to send your opinion on this issue to the letters to the editor column of the newspaper. Write your letter below. Say whether you agree or disagree with the argument you read in the newspaper and give reasons for your opinion.

Write in English in clear characters in the space below. Write around 100 words.

B. The distribution of ratings per rater

	2	3	4	5	6	7	8	9	10	Total
Austrian raters										
R1	1	2	6	8	9	7	11	14	2	60
R2	0	0	6	11	5	13	16	3	1	55
R3	0	0	2	8	23	13	8	3	3	60
R4	0	1	2	5	19	14	13	5	1	60
R5	0	1	4	3	18	12	12	4	6	60
R6	1	2	4	7	12	12	12	8	2	60
Finnish raters										
R7	6	53	36	5	0	NA	NA	NA	NA	100
R8	8	44	40	6	2	NA	NA	NA	NA	100
R9	8	46	44	2	0	NA	NA	NA	NA	100

Note: For the Austrian raters, the columns refer to the bands on the Austrian scale (2-10), so rater 1 awarded band 2 once, band 3 twice, band 4 six times etc. For the Finnish raters, the columns refer to the CEFR scores (2=A2, 3=B1, 4=B2, 5=C1, 6=C2), so rater 7 awarded A2 six times, B1 53 times, B2 36 times etc.

C. The number of common scripts rated by each pair of raters for Task 1

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6	Rater7	Rater8	Rater9
Rater1		20	20	20	10	20	30	30	30
Rater2	20		10	20	20	20	30	30	30
Rater3	20	10		20	20	20	30	30	30
Rater4	20	20	20		20	10	30	30	30
Rater5	10	20	20	20		20	30	30	30
Rater6	20	20	20	10	20		30	30	30
Rater7	30	30	30	30	30	30		50	50
Rater8	30	30	30	30	30	30	50		50
Rater9	30	30	30	30	30	30	50	50	

Note: The corresponding table for Task 2 is similar

D. The total number of ratings below B2/at B2 and above for Task 1 with outliers

	below B2	B2 and above	Total	Outliers low*	Outliers high*
Finnish holistic CEFR ratings	79 52.7 %	71 47.3 %	150 100 %	N.A.	N.A.
Austrian scale ratings (cut-off 24)	65 37.1 %	110 62.9 %	175 100 %	5/113 4.4 %	13/113 11.5 %
Austrian scale ratings (cut-off 25)	75 42.9 %	100 57.1 %	175 100 %	6/113 5.3 %	10/113 8.8 %
Austrian scale ratings (cut-off 26)	86 49.1 %	89 50.9 %	175 100 %	7/113 6.2 %	6/113 5.3 %
Austrian scale ratings (cut-off 27)	95 54.3 %	80 45.7 %	175 100 %	8/113 7.1 %	3/113 2.7 %
Austrian holistic CEFR ratings	69 39.4 %	106 60.6 %	175 100%	7/117 6.0 %	10/117 8.5 %

note: 5 ratings were missing for the Austrian raters

*Outliers refer to the number of Austrian ratings for candidates which were rated at B2 or above by all three Finnish raters but below B2 by the Austrian raters (Outliers low), and vice versa (Outliers high) (32 cases out of 50, whereby in 17 cases all Finnish raters awarded B1 and in 15 cases all of them awarded B2)

E. The total number of ratings below B2/at B2 and above for Task 2 with outliers

	B1 and below	B2 and above	Total	Outliers low*	Outliers high*
Finnish holistic CEFR ratings	86 57.3 %	64 42.7 %	150 100 %	N.A.	N.A.
Austrian scale ratings (cut-off 24)	59 32.8 %	121 67.2 %	180 100 %	1/96 1.0 %	16/96 16.7 %
Austrian scale ratings (cut-off 25)	71 39.4 %	109 60.6 %	180 100 %	2/96 2.1 %	14/96 14.6 %
Austrian scale ratings (cut-off 26)	79 43.9 %	101 56.1 %	180 100 %	3/96 3.1 %	14/96 14.6 %
Austrian scale ratings (cut-off 27)	89 49.4 %	91 50.6 %	180 100 %	5/96 5.2 %	11/96 11.5 %
Austrian holistic CEFR ratings	59 32.8 %	121 67.2 %	180 100 %	1/96 1.0 %	16/96 16.7 %

*Outliers refer to the number of Austrian ratings for candidates which were rated at B2 or above by all three Finnish raters but below B2 by the Austrian raters (Outliers low), and vice versa (Outliers high) (28 cases out of 50, whereby in 18 cases all Finnish raters awarded B1 and in 10 cases all of them awarded B2)

F. Task 1: Rater fit statistics (Analytic Austrian Scale, cut-off=24)

Rater		Outfit	Outfit_t	Outfit_p	Infit	Infit_t	Infit_p
1	Austrian	0.84	-0.07	0.94	0.77	-0.70	0.49
2		1.31	0.51	0.61	1.18	0.69	0.49
3		0.43	-0.36	0.72	0.69	-1.07	0.28
4		0.99	0.30	0.76	1.05	0.21	0.84
5		0.98	0.29	0.77	1.01	0.09	0.93
6		0.67	-0.21	0.84	0.85	-0.51	0.61
7	Finnish	0.87	0.02	0.99	1.01	0.05	0.96
8		0.95	0.05	0.96	1.01	0.05	0.96
9		0.87	-0.01	0.99	0.93	-0.37	0.71

Mean Rater Severity= -0.47 (SD=0.81); Mean Student Ability=0 (SD=2.56)

Expected A Priori (EAP) Reliability of student measures = 0.80

G. Task 2: Rater fit statistics (Analytic Austrian Scale, cut-off=24)

Rater		Outfit	Outfit_t	Outfit_p	Infit	Infit_t	Infit_p
1	Austrian	0.48	-0.15	0.88	0.75	-0.86	0.39
2		0.84	0.07	0.95	1.01	0.06	0.95
3		0.66	-0.33	0.74	0.84	-0.65	0.52
4		0.47	0.10	0.92	0.78	-0.68	0.50
5		1.93	0.85	0.39	1.22	0.78	0.44
6		1.05	0.30	0.76	1.07	0.28	0.78
7	Finnish	0.71	-0.22	0.82	0.92	-0.42	0.68
8		1.03	0.20	0.84	1.09	0.45	0.65
9		0.93	0.02	0.98	1.02	0.14	0.89

Mean Rater Severity= -0.59 (SD=1.07); Mean Student Ability=0 (SD=2.52)

Expected A Priori (EAP) Reliability of student measures = 0.79

H. Task 1: Rater fit statistics (Holistic Austrian scale)

Rater		Outfit	Outfit_t	Outfit_p	Infit	Infit_t	Infit_p
1	Austrian	0.94	0.08	0.94	0.83	-0.47	0.64
2		0.74	0.01	0.99	0.91	-0.27	0.79
3		0.50	-0.16	0.87	0.81	-0.61	0.54
4		0.76	0.04	0.96	0.91	-0.25	0.80
5		0.92	0.33	0.74	1.10	0.36	0.72
6		0.67	-0.11	0.92	0.87	-0.40	0.69
7	Finnish	0.85	0.03	0.97	1.01	0.06	0.96
8		0.98	0.14	0.89	1.06	0.32	0.75
9		0.95	0.08	0.94	0.92	-0.36	0.72

Mean Rater Severity= -0.46 (SD=0.68); Mean Student Ability=0 (SD=2.86)

Expected A Priori (EAP) Reliability of student measures = 0.80

I. Task 2: Rater fit statistics (Holistic Austrian scale)

Rater		Outfit	Outfit_t	Outfit_p	Infit	Infit_t	Infit_p
1	Austrian	0.45	-0.13	0.89	0.74	-0.90	0.37
2		0.87	0.10	0.92	1.03	0.11	0.91
3		0.68	-0.27	0.79	0.84	-0.63	0.53
4		0.48	0.13	0.90	0.78	-0.66	0.51
5		1.79	0.82	0.41	1.22	0.78	0.44
6		1.04	0.30	0.77	1.08	0.30	0.76
7	Finnish	0.72	-0.19	0.85	0.94	-0.31	0.76
8		1.05	0.22	0.83	1.10	0.50	0.61
9		0.93	0.03	0.98	1.02	0.09	0.93

Mean Rater Severity= -0.59 (SD=1.07); Mean Student Ability=0 (SD=2.27)

Expected A Priori (EAP) Reliability of student measures = 0.79