Assessing Writing 37 (2018) 57–67

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

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

Franz Holzknecht[a,*], Ari Huhta[b], Iasonas Lamprianou[c]

[a] *University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria*
[b] *University of Jyväskylä, Finland*
[c] *University of Cyprus, Cyprus*

ARTICLE INFO

Keywords:
CEFR levels
Assessing writing
Rating scales
Rater effects
Rasch model

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.

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 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

[*] Corresponding author.
E-mail address: franz.holzknecht@uibk.ac.at (F. Holzknecht).

https://doi.org/10.1016/j.asw.2018.03.009
Received 28 April 2017; Received in revised form 13 March 2018; Accepted 29 March 2018
1075-2935/ © 2018 Elsevier Inc. All rights reserved.

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 1222 texts, each rated by 5–10 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 0.700 and 0.882 (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.

Huhta, Alanen, Tarnanen, Martin, and Hirvela (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. Huhta et al. (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" (Huhta et al., 2014, p. 318).

Finally, in a study by Deygers and Van 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?

Download English Version:

https://daneshyari.com/en/article/6831490

Download Persian Version:

https://daneshyari.com/article/6831490

Daneshyari.com